Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)*

Andrew J. Patton, Duke University
Johanna F. Ziegel, University of Bern
Rui Chen, Duke University

First version: 5 December 2015. This version: 18 December 2018.

Abstract

Expected Shortfall (ES) is the average return on a risky asset conditional on the return being below some quantile of its distribution, namely its Value-at-Risk (VaR). The Basel III Accord, which will be implemented in the years leading up to 2019, places new attention on ES, but unlike VaR, there is little existing work on modeling ES. We use recent results from statistical decision theory to overcome the problem of "elicitability" for ES by *jointly* modelling ES and VaR, and propose new dynamic models for these risk measures. We provide estimation and inference methods for the proposed models, and confirm via simulation studies that the methods have good finite-sample properties. We apply these models to daily returns on four international equity indices, and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models.

Keywords: Risk management, tails, crashes, forecasting, generalized autoregressive score.
J.E.L. codes: G17, C22, G32, C58.

*For helpful comments we thank Tim Bollerslev, Timo Dimitriadis, Rob Engle, Tobias Fissler, Jia Li, Nour Meddahi, and seminar participants at the Bank of Japan, Cambridge University, Deutsche Bundesbank, Duke University, EPFL, Federal Reserve Bank of New York, Hitotsubashi University, New York University, Toulouse School of Economics, University of Illinois-Urbana Champaign, University of Southern California, University of Tennessee, University of Western Ontario, the 2017 EC^2 conference in Amsterdam, and the 2015 Oberwolfach Workshop on Quantitative Risk Management where this project started. The first author would particularly like to thank the finance department at NYU Stern, where much of his work on this paper was completed. A MATLAB toolbox for this article is available at `www.econ.duke.edu/~ap172/research.html`. Contact address: Andrew Patton, Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097. Email: `andrew.patton@duke.edu`.

1 Introduction

The financial crisis of 2007-08 and its aftermath led to numerous changes in financial market regulation and banking supervision. One important change appears in the Third Basel Accord (Basel Committee, 2010), where new emphasis is placed on "Expected Shortfall" (ES) as a measure of risk, complementing, and in parts substituting, the more-familiar Value-at-Risk (VaR) measure. Expected Shortfall is the expected return on an asset conditional on the return being below a given quantile of its distribution, namely its VaR. That is, if Y_t is the return on some asset over some horizon (e.g., one day or one week) with conditional (on information set $\mathcal{F}_{t-1}$) distribution F_t, which we assume to be strictly increasing with finite mean, the α-level VaR and ES are:

$$\text{ES}_t = \mathbb{E}\left[Y_t | Y_t \leq \text{VaR}_t, \mathcal{F}_{t-1}\right] \tag{1}$$

$$\text{where} \quad \text{VaR}_t = F_t^{-1}(\alpha), \text{ for } \alpha \in (0,1) \tag{2}$$

$$\text{and} \quad Y_t | \mathcal{F}_{t-1} \sim F_t \tag{3}$$

As Basel III is implemented worldwide (implementation is expected to occur in the period leading up to January 1st, 2019), ES will inevitably gain, and require, increasing attention from risk managers and banking supervisors and regulators. The new "market discipline" aspects of Basel III mean that ES and VaR will be regularly disclosed by banks, and so a knowledge of these measures will also likely be of interest to these banks' investors and counter-parties.

There is, however, a paucity of empirical models for expected shortfall. The large literature on volatility models (see Andersen *et al.* (2006) for a review) and VaR models (see Komunjer (2013) and McNeil et al. (2015)), have provided many useful models for these measures of risk. However, while ES has long been known to be a "coherent" measure of risk (Artzner, *et al.* 1999), in contrast with VaR, the literature contains relatively few models for ES; some exceptions are discussed below. This dearth is perhaps in part because regulatory interest in this risk measure is only recent, and may also be due to the fact that this measure is not "elicitable." A risk measure (or statistical functional more generally) is said to be "elicitable" if there exists a loss function such that the risk measure is the solution to minimizing the expected loss. For example, the mean is elicitable using the quadratic loss function, and VaR is elicitable using the piecewise-linear or "tick" loss function.

Having such a loss function is a stepping stone to building dynamic models for these quantities. We use recent results from Fissler and Ziegel (2016), who show that ES is *jointly elicitable* with VaR, to build new dynamic models for ES and VaR.

This paper makes three main contributions. Firstly, we present some novel dynamic models for ES and VaR, drawing on the GAS framework of Creal, *et al.* (2013), as well as successful models from the volatility literature, see Andersen *et al.* (2006). The models we propose are semiparametric in that they impose parametric structures for the dynamics of ES and VaR, but are completely agnostic about the conditional distribution of returns (aside from regularity conditions required for estimation and inference). The models proposed in this paper are related to the class of "CAViaR" models proposed by Engle and Manganelli (2004a), in that we directly parameterize the measure(s) of risk that are of interest, and avoid the need to specify a conditional distribution for returns. The models we consider make estimation and prediction fast and simple to implement. Our semiparametric approach eliminates the need to specify and estimate a conditional density, thereby removing the possibility that such a model is misspecified, though at a cost of a loss of efficiency compared with a correctly specified density model.

Our second contribution is asymptotic theory for a general class of dynamic semiparametric models for ES and VaR. This theory is an extension of results for VaR presented in Weiss (1991) and Engle and Manganelli (2004a), and draws on identification results in Fissler and Ziegel (2016) and results for M-estimators in Newey and McFadden (1994). We present conditions under which the estimated parameters of the VaR and ES models are consistent and asymptotically normal, and we present a consistent estimator of the asymptotic covariance matrix. We show via an extensive Monte Carlo study that the asymptotic results provide reasonable approximations in realistic simulation designs. In addition to being useful for the new models we propose, the asymptotic theory we present provides a general framework for other researchers to develop, estimate, and evaluate new models for VaR and ES.

Our third contribution is an extensive application of our new models and estimation methods in an out-of-sample analysis of forecasts of ES and VaR for four international equity indices over the period January 1990 to December 2016. We compare these new models with existing methods

from the literature across a range of tail probability values (α) used in risk management. We use Diebold and Mariano (1995) tests to identify the best-performing models for ES and VaR, and we present simple regression-based methods, related to those of Engle and Manganelli (2004a) and Nolde and Ziegel (2017), to "backtest" the ES forecasts.

Some work on expected shortfall estimation and prediction has appeared in the literature, overcoming the problem of elicitability in different ways: Engle and Manganelli (2004b) discuss using extreme value theory, combined with GARCH or CAViaR dynamics, to obtain forecasts of ES. Cai and Wang (2008) propose estimating VaR and ES based on nonparametric conditional distributions, while Taylor (2008) and Gschöpf *et al.* (2015) estimate models for "expectiles" (Newey and Powell, 1987) and map these to ES. Zhu and Galbraith (2011) propose using flexible parametric distributions for the standardized residuals from models for the conditional mean and variance. Drawing on Fissler and Ziegel (2016), we overcome the problem of elicitability more directly, and open up new directions for ES modeling and prediction.

In recent independent work, Taylor (2017) proposes using the asymmetric Laplace distribution to jointly estimate dynamic models for VaR and ES. He shows the intriguing result that the negative log-likelihood of this distribution corresponds to one of the loss functions presented in Fissler and Ziegel (2016), and thus can be used to estimate and evaluate such models. Unlike our paper, Taylor (2017) provides no asymptotic theory for his proposed estimation method, nor any simulation studies of its reliability. However, given the link he presents, the theoretical results we present below can be used to justify *ex post* the methods of his paper.

The remainder of the paper is structured as follows. In Section 2 we present new dynamic semiparametric models for ES and VaR and compare them with the main existing models for ES and VaR. In Section 3 we present asymptotic distribution theory for a generic dynamic semiparametric model for ES and VaR, and in Section 4 we study the finite-sample properties of the estimators in some realistic Monte Carlo designs. In Section 5 we apply the new models to daily data on four international equity indices, and compare these models both in-sample and out-of-sample with existing models. Section 6 concludes. Proofs and additional technical details are presented in the appendix, and a supplemental web appendix contains detailed proofs and additional analyses.

2 Dynamic models for ES and VaR

In this section we propose some new dynamic models for expected shortfall (ES) and Value-at-Risk (VaR). We do so by exploiting recent work in Fissler and Ziegel (2016) which shows that these variables are elicitable *jointly*, despite the fact that ES was known to be not elicitable on its own, see Gneiting (2011a). The models we propose are based on the GAS framework of Creal, *et al.* (2013) and Harvey (2013), which we briefly review in Section 2.2 below.

2.1 A consistent scoring rule for ES and VaR

Fissler and Ziegel (2016) show that the following class of loss functions (or "scoring rules"), indexed by the functions G_1 and G_2, is consistent for VaR and ES. That is, minimizing the expected loss using any of these loss functions returns the true VaR and ES. In the functions below, we use the notation v and e for VaR and ES.

$$\begin{aligned} L_{FZ}\left(Y,v,e;\alpha,G_1,G_2\right) &= \left(\mathbf{1}\left\{Y\leq v\right\}-\alpha\right)\left(G_1\left(v\right)-G_1\left(Y\right)+\frac{1}{\alpha}G_2\left(e\right)v\right) \\ &\quad -G_2\left(e\right)\left(\frac{1}{\alpha}\mathbf{1}\left\{Y\leq v\right\}Y-e\right)-\mathcal{G}_2\left(e\right) \end{aligned} \tag{4}$$

where G_1 is weakly increasing, G_2 is strictly increasing and strictly positive, and $\mathcal{G}_2' = G_2$. We will refer to the above class as "FZ loss functions."[1] Minimizing any member of this class yields VaR and ES:

$$\left(\text{VaR}_t, \text{ES}_t\right) = \arg\min_{(v,e)} \mathbb{E}_{t-1}\left[L_{FZ}\left(Y_t,v,e;\alpha,G_1,G_2\right)\right] \tag{5}$$

Using the FZ loss function for estimation and forecast evaluation requires choosing G_1 and G_2. To do so, first define $\Delta L\left(Y_t,v_{1t},e_{1t},v_{2t},e_{2t}\right) \equiv L\left(Y_t,v_{1t},e_{1t}\right) - L\left(Y_t,v_{2t},e_{2t}\right)$ as the loss difference for two forecasts $(v_{j,t},e_{j,t})$, $j\in\{1,2\}$. We choose G_1 and G_2 so that the loss function generates ΔL that is homogeneous of degree zero, a property that has been shown in volatility forecasting applications to lead to higher power in Diebold-Mariano (1995) tests, see Patton and Sheppard (2009). Nolde and Ziegel (2017) show that there does *not* generally exist an FZ loss function that

[1] Consistency of the FZ loss function for VaR and ES also requires imposing that $e \leq v$, which follows naturally from the definitions of ES and VaR in equations (1) and (2). We discuss how we impose this restriction empirically in Sections 4 and 5 below.

generates loss differences that are homogeneous of degree zero, however, we show in Proposition 1 below that zero-degree homogeneity may be attained by exploiting the fact that, for the values of α that are of interest in risk management applications (namely, values ranging from around 0.01 to 0.10), we may assume that $\text{ES}_t < 0$ a.s. $\forall\ t$. Proposition 1 shows that if we further impose that $\text{VaR}_t < 0$ a.s. $\forall\ t$, then, up to irrelevant location and scale factors, there is only *one* FZ loss function that generates loss differences that are homogeneous of degree zero.[2] The uniqueness of the loss function defined in Proposition 1 means, of course, that it also has the added benefit of there being no remaining shape or tuning parameters to be specified.

Proposition 1 *Define the FZ loss difference for two forecasts* (v_{1t}, e_{1t}) *and* (v_{2t}, e_{2t}) *as* $L_{FZ}\left(Y_t, v_{1t}, e_{1t}; \alpha, G_1, G_2\right) - L_{FZ}\left(Y_t, v_{2t}, e_{2t}; \alpha, G_1, G_2\right)$. *Under the assumption that VaR and ES are both strictly negative, the loss differences generated by a FZ loss function are homogeneous of degree zero iff* $G_1(x) = 0$ *and* $G_2(x) = -1/x$. *The resulting "FZ0" loss function is:*

$$L_{FZ0}\left(Y, v, e; \alpha\right) = -\frac{1}{\alpha e}\mathbf{1}\left\{Y \leq v\right\}\left(v - Y\right) + \frac{v}{e} + \log\left(-e\right) - 1 \tag{6}$$

All proofs are presented in Appendix A. In Figure 1 we plot L_{FZ0} when $Y = -1$. In the left panel we fix $e = -2.06$ and vary v, and in the right panel we fix $v = -1.64$ and vary e. (These values for (v, e) are the $\alpha = 0.05$ VaR and ES from a standard Normal distribution.) The left panel shows that the implied VaR loss function resembles the "tick" loss function from quantile estimation, see Komunjer (2005) for example. In the right panel we see that the implied ES loss function resembles the "QLIKE" loss function from volatility forecasting, see Patton (2011) for example. In both panels, values of (v, e) where $v < e$ are presented with a dashed line, as by definition ES_t is below VaR_t, and so such values that would never be considered in practice. In Figure 2 we plot the contours of expected FZ0 loss for a standard Normal random variable. The minimum value, which is attained when $(v, e) = (-1.64, -2.06)$, is marked with a star, and we see that the

[2] If VaR can be positive, then there is one free shape parameter in the class of zero-homogeneous FZ loss functions (φ_1/φ_2, in the notation of the proof of Proposition 1). In that case, our use of the loss function in equation (6) can be interpreted as setting that shape parameter to zero. This shape parameter does not affect the consistency of the loss function, as it is a member of the FZ class, but it may affect the ranking of misspecified models, see Patton (2016).

"iso-expected loss" contours (that is, the level sets) of the expected loss function are boundaries of convex sets. Fissler (2017) shows that convexity of sublevel sets holds more generally for the FZ0 loss function under any distribution with finite first moments, unique α-quantiles, continuous densities, and negative ES.

[INSERT FIGURES 1 AND 2 ABOUT HERE]

With the FZ0 loss function in hand, it is then possible to consider semiparametric dynamic models for ES and VaR:

$$(\text{VaR}_t, \text{ES}_t) = (v(\mathbf{Z}_{t-1}; \boldsymbol{\theta}), e(\mathbf{Z}_{t-1}; \boldsymbol{\theta})) \tag{7}$$

that is, where the true VaR and ES are some specified parametric functions of elements of the information set, $\mathbf{Z}_{t-1} \in \mathcal{F}_{t-1}$. The parameters of this model are estimated via:

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^{T} L_{FZ0}(Y_t, v(\mathbf{Z}_{t-1}; \boldsymbol{\theta}), e(\mathbf{Z}_{t-1}; \boldsymbol{\theta}); \alpha) \tag{8}$$

Such models impose a parametric structure on the dynamics of VaR and ES, through their relationship with lagged information, but require no assumptions, beyond regularity conditions, on the conditional distribution of returns. In this sense, these models are semiparametric. Using theory for M-estimators (see White (1994) and Newey and McFadden (1994) for example) we establish in Section 3 below the asymptotic properties of such estimators. Before doing so, we first consider some new dynamic specifications for ES and VaR.

2.2 A GAS model for ES and VaR

One of the challenges in specifying a dynamic model for a risk measure, or any other quantity of interest, is the mapping from lagged information to the current value of the variable. Our first proposed specification for ES and VaR draws on the work of Creal, *et al.* (2013) and Harvey (2013), who proposed a general class of models called "generalized autoregressive score" (GAS) models by the former authors, and "dynamic conditional score" models by the latter author. In both cases the models start from an assumption that the target variable has some parametric conditional distribution, where the parameter (vector) of that distribution follows a GARCH-like equation.

The forcing variable in the model is the lagged score of the log-likelihood, scaled by some positive definite matrix, a common choice for which is the inverse Hessian. This specification nests many well known models, including ARMA, GARCH (Bollerslev, 1986) and ACD (Engle and Russell, 1998) models. See Koopman *et al.* (2016) for an overview of GAS and related models.

We adopt this modeling approach and apply it to our M-estimation problem. In this application, the forcing variable is a function of the derivative and Hessian of the L_{FZ0} loss function rather than a log-likelihood. We will consider the following GAS(1,1) model for ES and VaR:

$$\begin{bmatrix} v_{t+1} \\ e_{t+1} \end{bmatrix} = \mathbf{w} + \mathbf{B} \begin{bmatrix} v_t \\ e_t \end{bmatrix} + \mathbf{A}\mathbf{H}_t^{-1}\boldsymbol{\nabla}_t \tag{9}$$

where $\mathbf{w}$ is a (2×1) vector and $\mathbf{B}$ and $\mathbf{A}$ are (2×2) matrices. The forcing variable in this specification is comprised of two components, $\mathbf{H}_t$ and $\boldsymbol{\nabla}_t$. Using details provided in Appendix B.1, the latter can be shown to be:

$$\boldsymbol{\nabla}_t \equiv \begin{bmatrix} \partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)/\partial v_t \\ \partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)/\partial e_t \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha v_t e_t}\lambda_{v,t} \\ \frac{-1}{\alpha e_t^2}\left(\lambda_{v,t} + \alpha\lambda_{e,t}\right) \end{bmatrix} \tag{10}$$

$$\text{where} \quad \lambda_{v,t} \equiv -v_t\left(\mathbf{1}\left\{Y_t \leq v_t\right\} - \alpha\right) \tag{11}$$

$$\lambda_{e,t} \equiv \frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t\right\} Y_t - e_t \tag{12}$$

Note that the expression given for $\partial L_{FZ0}/\partial v_t$ only holds for $Y_t \neq v_t$. As we assume that Y_t is continuously distributed, this holds with probability one. The scaling matrix, $\mathbf{H}_t$, is related to the Hessian:

$$\mathbf{I}_t \equiv \begin{bmatrix} \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t^2} & \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t \partial e_t} \\ \bullet & \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial e_t^2} \end{bmatrix} = \begin{bmatrix} -\frac{f_t(v_t)}{\alpha e_t} & 0 \\ 0 & \frac{1}{e_t^2} \end{bmatrix} \tag{13}$$

The second equality above exploits the fact that $\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]/\partial v_t \partial e_t = 0$ under the assumption that the dynamics for VaR and ES are correctly specified. The first element of the matrix $\mathbf{I}_t$ depends on the unknown conditional density of Y_t. We would like to avoid estimating this density, and we approximate the term $f_t\left(v_t\right)$ as being proportional to v_t^{-1}. This approximation holds exactly if Y_t is a zero-mean location-scale random variable, $Y_t = \sigma_t \eta_t$, where $\eta_t \sim iid\ F_\eta\left(0, 1\right)$, as in that case we have:

$$f_t\left(v_t\right) = f_t\left(\sigma_t v_\alpha\right) = \frac{1}{\sigma_t} f_\eta\left(v_\alpha\right) \equiv k_\alpha \frac{1}{v_t} \tag{14}$$

where $k_\alpha \equiv v_\alpha f_\eta(v_\alpha)$ is a constant with the same sign as v_t. We define $\mathbf{H}_t$ to equal $\mathbf{I}_t$ with the first element replaced using the approximation in the above equation.[3] The forcing variable in our GAS model for VaR and ES then becomes:

$$\mathbf{H}_t^{-1}\boldsymbol{\nabla}_t = \begin{bmatrix} \frac{-1}{k_\alpha}\lambda_{v,t} \\ \frac{-1}{\alpha}\left(\lambda_{v,t} + \alpha\lambda_{e,t}\right) \end{bmatrix} \tag{15}$$

Notice that the second term in the model is a linear combination of the two elements of the forcing variable, and since the forcing variable is premultiplied by a coefficient matrix, say $\tilde{\mathbf{A}}$, we can equivalently use

$$\begin{aligned} \tilde{\mathbf{A}}\mathbf{H}_t^{-1}\boldsymbol{\nabla}_t &= \mathbf{A}\boldsymbol{\lambda}_t \\ \text{where } \boldsymbol{\lambda}_t &\equiv [\lambda_{v,t}, \lambda_{e,t}]' \end{aligned} \tag{16}$$

We choose to work with the $\mathbf{A}\boldsymbol{\lambda}_t$ parameterization, as the two elements of this forcing variable $(\lambda_{v,t}, \lambda_{e,t})$ are not directly correlated, while the elements of $\mathbf{H}_t^{-1}\boldsymbol{\nabla}_t$ are correlated due to the overlapping term $(\lambda_{v,t})$ appearing in both elements. This aids the interpretation of the results of the model without changing its fit.

To gain some intuition for how past returns affect current forecasts of ES and VaR in this model, consider the "news impact curve" of this model, which presents (v_{t+1}, e_{t+1}) as a function of Y_t through its impact on $\boldsymbol{\lambda}_t \equiv [\lambda_{v,t}, \lambda_{e,t}]'$, holding all other variables constant. Figure 3 shows these two curves for $\alpha = 0.05$, using the estimated parameters for this model when applied to daily returns on the S&P 500 index (details are presented in Section 5 below). We consider two values for the "current" value of (v, e): 10% above and below the long-run average for these variables. We see that for values where $Y_t > v_t$, the news impact curves are flat, reflecting the fact that on those days the value of the realized return does not enter the forcing variable. When $Y_t \leq v_t$, we see that ES and VaR react linearly to Y and this reaction is through the $\lambda_{e,t}$ forcing variable; the reaction through the $\lambda_{v,t}$ forcing variable is a simple step (down) in both of these risk measures.

[3] Note that we do *not* use the fact that the scaling matrix is exactly the inverse Hessian (e.g., by invoking the information matrix equality) in our empirical application or our theoretical analysis. Also, note that if we considered a value of α for which $v_t = 0$, then $v_\alpha = 0$ and we cannot justify our approximation using this approach. However, we focus on cases where $\alpha \ll 1/2$, and so we are comfortable assuming $v_t \neq 0$, making k_α invertible.

[INSERT FIGURE 3 ABOUT HERE]

2.3 A one-factor GAS model for ES and VaR

The specification in Section 2.2 allows ES and VaR to evolve as two separate, correlated, processes. In many risk forecasting applications, a useful simpler model is one based on a structure with only one time-varying risk measure, e.g. volatility. We will consider a one-factor model in this section, and will name the model in Section 2.2 a "two-factor" GAS model.

Consider the following one-factor GAS model for ES and VaR, where both risk measures are driven by a single variable, κ_t.[4]

$$\begin{aligned} v_t &= a\exp\{\kappa_t\} \\ e_t &= b\exp\{\kappa_t\}, \text{ where } b<a<0 \\ \text{and } \kappa_t &= \omega+\beta\kappa_{t-1}+\gamma H_{t-1}^{-1}s_{t-1} \end{aligned} \tag{17}$$

The forcing variable, $H_{t-1}^{-1}s_{t-1}$, in the evolution equation for κ_t is obtained from the FZ0 loss function, plugging in $(a\exp\{\kappa_t\}, b\exp\{\kappa_t\})$ for (v_t, e_t). Using details provided in Appendix B.2, we find that the score and Hessian are:

$$s_t \equiv \frac{\partial L_{FZ0}\left(Y_t, a\exp\{\kappa_t\}, b\exp\{\kappa_t\};\alpha\right)}{\partial\kappa} = -\frac{1}{e_t}\left(\frac{1}{\alpha}\mathbf{1}\{Y_t\le v_t\}Y_t - e_t\right) \tag{18}$$

$$\text{and } I_t \equiv \frac{\partial^2\mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, a\exp\{\kappa_t\}, b\exp\{\kappa_t\};\alpha\right)\right]}{\partial\kappa_t^2} = \frac{\alpha-k_\alpha a_\alpha}{\alpha} \tag{19}$$

where k_α is a negative constant and a_α lies between zero and one. The Hessian, I_t, turns out to be a constant in this case, and since we estimate a free coefficient on our forcing variable, we can set the scaling matrix, H_t, to any positive constant; we set H_t to one. Note that the VaR score, $\lambda_{v,t} = \partial L/\partial v$, turns out to drop out from the forcing variable. Thus the one-factor GAS model for

[4] We use the structure in equation (17) to emphasize its similarity to conditional volatility models, which we include as competitor models in the next section. The one-factor model for ES and VaR can also be obtained by considering a zero-mean volatility model for Y_t, with *iid* standardized residuals, say denoted η_t. In this case, κ_t is the log conditional standard deviation of Y_t, and $a = F_\eta^{-1}(\alpha)$ and $b = \mathbb{E}[\eta|\eta\le a]$. (We exploit this interpretation when linking these models to GARCH models in Section 2.5.1 below.)

ES and VaR becomes:

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma \frac{1}{b\exp\{\kappa_{t-1}\}}\left(\frac{1}{\alpha}\mathbf{1}\{Y_{t-1} \leq a\exp\{\kappa_{t-1}\}\}Y_{t-1} - b\exp\{\kappa_{t-1}\}\right) \tag{20}$$

We drop the negative sign in s_t that its coefficient, γ, is positive rather than negative. This change, of course, does not affect the fit of the model. The FZ loss function only identifies (v_t, e_t), and in the specification in equation (17) this implies that ω, a, and b are not separably identifiable: for any constant c, the parameter vectors $(\omega, a, b, \beta, \gamma)$ and $(w + c(1-\beta), a\exp\{-c\}, b\exp\{-c\}, \beta, \gamma)$ yield identical sequences of (v_t, e_t), and thus identical values of the objective function. Fixing any one of ω, a, or b resolves this problem; we set $\omega = 0$ for simplicity.

Foreshadowing the empirical results in Section 5, we find that this one-factor GAS model outperforms the two-factor GAS model in out-of-sample forecasts for most of the asset return series that we study.

2.4 Existing dynamic models for ES and VaR

As noted in the introduction, there is a relative paucity of dynamic models for ES and VaR, but there is not a complete absence of such models. The simplest existing model is based on a rolling window estimate of these quantities:

$$\begin{aligned} \widehat{\mathrm{VaR}}_t &= \widehat{Quantile}\{Y_s\}_{s=t-m}^{t-1} \\ \widehat{\mathrm{ES}}_t &= \frac{1}{\alpha m}\sum_{s=t-m}^{t-1} Y_s \mathbf{1}\left\{Y_s \leq \widehat{\mathrm{VaR}}_s\right\} \end{aligned} \tag{21}$$

where $\widehat{Quantile}\{Y_s\}_{s=t-m}^{t-1}$ denotes the sample quantile of Y_s over the period $s \in [t-m, t-1]$. Common choices for the window size, m, include 125, 250 and 500, corresponding to six months, one year and two years of daily return observations respectively.

A more challenging competitor for the new ES and VaR models proposed in this paper are those based on ARMA-GARCH dynamics for the conditional mean and variance, accompanied by some assumption for the distribution of the standardized residuals. These models all take the form:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \eta_t \\ \eta_t &\sim iid\ F_\eta(0,1) \end{aligned} \tag{22}$$

where μ_t and σ_t^2 are specified to follow some ARMA and GARCH model, and $F_\eta(0,1)$ is some arbitrary, strictly increasing, distribution with mean zero and variance one. What remains is to specify a distribution for the standardized residual, η_t. Given a choice for F_η, VaR and ES forecasts are obtained as:

$$\begin{aligned} v_t &= \mu_t + a\sigma_t, \quad \text{where} \quad a = F_\eta^{-1}(\alpha) \\ e_t &= \mu_t + b\sigma_t, \quad \text{where} \quad b = \mathbb{E}[\eta_t | \eta_t \leq a] \end{aligned} \tag{23}$$

Two parametric choices for F_η are common in the literature:

$$\begin{aligned} \eta_t &\sim iid\ N(0,1) \\ \eta_t &\sim iid\ Skew\ t(0,1,\nu,\lambda) \end{aligned} \tag{24}$$

There are various skew t distributions used in the literature; in the empirical analysis below we use that of Hansen (1994). A nonparametric alternative is to estimate the distribution of η_t using the empirical distribution function (EDF), an approach that is also known as "filtered historical simulation," and one that is perhaps the best existing model for ES, see the survey by Engle and Manganelli (2004b).[5] We consider all of these models in our empirical analysis in Section 5.

2.5 GARCH and ES/VaR estimation

In this section we consider two extensions of the models presented above, in an attempt to combine the success and parsimony of GARCH models with this paper's focus on ES and VaR forecasting.

2.5.1 Estimating a GARCH model via FZ minimization

If an ARMA-GARCH model, including the specification for the distribution of standardized residuals, is correctly specified for the conditional distribution of an asset return, then maximum likelihood is the most efficient estimation method, and should naturally be adopted. If, on the other hand, we

[5] Some authors have also considered modeling the tail of F_η using extreme value theory, however for the relatively non-extreme values of α we consider here, past work (e.g., Engle and Manganelli (2004b), Nolde and Ziegel (2016) and Taylor (2017)) has found EVT to perform no better than the EDF, and so we do not include it in our analysis.

consider an ARMA-GARCH model only as a useful approximation to the true conditional distribution, then it is no longer clear that MLE is optimal. In particular, if the application of the model is to ES and VaR forecasting, then we might be able to improve the fitted ARMA-GARCH model by estimating the parameters of that model via FZ loss minimization, as discussed in Section 2.1. This estimation method is related to one discussed in Remark 1 of Francq and Zakoïan (2015).

Consider the following model for asset returns:

$$\begin{aligned} Y_t &= \sigma_t \eta_t, \quad \eta_t \sim iid\ F_\eta(0,1) \\ \sigma_t^2 &= \omega + \beta \sigma_{t-1}^2 + \gamma Y_{t-1}^2 \end{aligned} \tag{25}$$

The variable σ_t^2 is the conditional variance and is assumed to follow a GARCH(1,1) process. This model implies a structure analogous to the one-factor GAS model presented in Section 2.3, as we find:

$$\begin{aligned} v_t &= a \cdot \sigma_t, \ \text{ where } a = F_\eta^{-1}(\alpha) \\ e_t &= b \cdot \sigma_t, \ \text{ where } b = \mathbb{E}[\eta | \eta \leq a] \end{aligned} \tag{26}$$

Some further results on VaR and ES in dynamic location-scale models are presented in Appendix B.3. To apply this model to VaR and ES forecasting, we also have to estimate the VaR and ES of the standardized residual, denoted (a, b). Rather than estimating the parameters of this model using (Q)MLE, we consider here estimating via FZ loss minimization. As in the one-factor GAS model, ω is unidentified and we set it to one,[6] so the parameter vector to be estimated is (β, γ, a, b). This estimation approach leads to a fitted GARCH model that is tailored to provide the best-fitting ES and VaR forecasts, rather than the best-fitting volatility forecasts.

[6] Similar to the one-factor GAS model, in this case we find that for any strictly positive constant c, the parameter vectors $(\omega, a, b, \beta, \gamma)$ and $(cw, a/\sqrt{c}, b/\sqrt{c}, \beta, c\gamma)$ yield identical sequences of (v_t, e_t), and thus identical values of the objective function. Fixing any one of ω, a, or b resolves this problem. As ω must be strictly positive in a GARCH model, we cannot set it to zero as we did for the one-factor GAS model; instead we set it to one.

2.5.2 A hybrid GAS/GARCH model

Finally, we consider a direct combination of the forcing variable suggested by a GAS structure for a one-factor model of returns, described in equation (20), with the successful GARCH model for volatility. We specify:

$$\begin{aligned} Y_t &= \exp\{\kappa_t\}\eta_t, \quad \eta_t \sim iid\ F_\eta(0,1) \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma\frac{1}{e_{t-1}}\left(\frac{1}{\alpha}\mathbf{1}\{Y_{t-1} \leq v_{t-1}\}Y_{t-1} - e_{t-1}\right) + \delta\log|Y_{t-1}| \end{aligned} \tag{27}$$

The variable κ_t is the log-volatility, identified up to scale. As the latent variable in this model is log-volatility, we use the lagged log absolute return rather than the lagged squared return, so that the units remain in line for the evolution equation for κ_t. There are five parameters in this model $(\beta, \gamma, \delta, a, b)$, and we estimate them using FZ loss minimization.

3 Estimation of dynamic models for ES and VaR

This section presents asymptotic theory for the estimation of dynamic ES and VaR models by minimizing FZ loss. Given a sample of observations $(Y_1, \cdots, Y_T)$ and a constant $\alpha \in (0, 0.5)$, we are interested in estimating and forecasting the conditional α quantile (VaR) and corresponding expected shortfall (ES) of Y_t. Suppose Y_t is a real-valued random variable that has, conditional on information set $\mathcal{F}_{t-1}$, distribution function $F_t(\cdot|\mathcal{F}_{t-1})$ and corresponding density function $f_t(\cdot|\mathcal{F}_{t-1})$. Let $v_1(\boldsymbol{\theta}^0)$ and $e_1(\boldsymbol{\theta}^0)$ be some initial conditions for VaR and ES and let $\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, \mathbf{X}_{t-1}, \cdots, Y_1, \mathbf{X}_1\}$, where $\mathbf{X}_t$ is a vector of exogenous variables or predetermined variables, be the information set available for forecasting Y_t. The vector of unknown parameters to be estimated is $\boldsymbol{\theta}^0 \in \Theta \subset \mathbb{R}^p$.

The conditional VaR and ES of Y_t at probability level α, that is $\text{VaR}_\alpha(Y_t|\mathcal{F}_{t-1})$ and $\text{ES}_\alpha(Y_t|\mathcal{F}_{t-1})$, are assumed to follow some dynamic model:

$$\begin{bmatrix} \text{VaR}_\alpha(Y_t|\mathcal{F}_{t-1}) \\ \text{ES}_\alpha(Y_t|\mathcal{F}_{t-1}) \end{bmatrix} = \begin{bmatrix} v(Y_{t-1}, \mathbf{X}_{t-1}, \cdots, Y_1, \mathbf{X}_1; \boldsymbol{\theta}^0) \\ e(Y_{t-1}, \mathbf{X}_{t-1}, \cdots, Y_1, \mathbf{X}_1; \boldsymbol{\theta}^0) \end{bmatrix} \equiv \begin{bmatrix} v_t(\boldsymbol{\theta}^0) \\ e_t(\boldsymbol{\theta}^0) \end{bmatrix}, \ t = 1, \cdots, T. \tag{28}$$

The unknown parameters are estimated as:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_T &\equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \; L_T(\boldsymbol{\theta}) \\ \text{where} \;\; L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T} L_{FZ0}\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right); \alpha\right) \end{aligned} \tag{29}$$

and the FZ loss function L_{FZ0} is defined in equation (6). Below we provide conditions under which estimation of these parameters via FZ loss minimization leads to a consistent and asymptotically normal estimator, with standard errors that can be consistently estimated. In Supplemental Appendix SA.2 we show that all of these conditions are satisfied for the widely-used GARCH(1,1) model, drawing on Lumsdaine (1996) and Carrasco and Chen (2002) among others. See Francq and Zakoïan (2010) for a review of asymptotic theory for GARCH processes.

Assumption 1 *(A)* $L\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right); \alpha\right)$ *obeys the uniform law of large numbers.*

(B)(i) Θ *is a compact subset of* $\mathbb{R}^p$ *for* $p < \infty$*. (ii)*$\{Y_t\}_{t=1}^{\infty}$ *is a strictly stationary process. Conditional on all the past information* $\mathcal{F}_{t-1}$*, the distribution of* Y_t *is* $F_t\left(\cdot|\mathcal{F}_{t-1}\right)$ *which, for all* t*, belongs to a class of distribution functions on* $\mathbb{R}$ *with finite first moments and unique* α*-quantiles. (iii)* $\forall t$*, both* $v_t(\boldsymbol{\theta})$ *and* $e_t(\boldsymbol{\theta})$ *are* $\mathcal{F}_{t-1}$*-measurable and a.s. continuous in* $\boldsymbol{\theta}$*. (iv) If* $\Pr\left[v_t(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}^0) \cap e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0)\right] = 1 \; \forall \; t$*, then* $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Theorem 1 (Consistency) *Under Assumption 1,* $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}^0$ *as* $T \to \infty$.

The proof of Theorem 1, provided in Appendix A, is straightforward given Theorem 2.1 of Newey and McFadden (1994) and Corollary 5.5 of Fissler and Ziegel (2016). Assumption 1(A) can be satisfied by one of a variety of uniform laws of large numbers for the time series applications we consider here, see Andrews (1987) and Pötscher and Prucha (1989) for example. Assumption 1(B) is standard for parameter time series inference. Zwingmann and Holzmann (2016) show that if the α-quantile is not unique (violating part of our Assumption 1(B)(ii)), then the convergence rate and asymptotic distribution of $(\hat{v}_T, \hat{e}_T)$ are non-standard, even in a setting with *iid* data. We do not consider such problematic cases here.

We next turn to the asymptotic distribution of our parameter estimator. In the assumptions

below, K denotes a finite constant that can change from line to line, and we use $\|\mathbf{x}\|$ to denote the Euclidean norm of if $\mathbf{x}$ is a vector, and the Frobenius norm if $\mathbf{x}$ is a matrix.

Assumption 2 *(A) For all t, we have (i) $v_t(\boldsymbol{\theta})$ and $e_t(\boldsymbol{\theta})$ are a.s. twice continuously differentiable in $\boldsymbol{\theta}$, (ii) $e_t\left(\boldsymbol{\theta}^0\right) < v_t(\boldsymbol{\theta}^0) \leq 0$.*

(B) For all t, we have (i) conditional on all the past information $\mathcal{F}_{t-1}$, Y_t has a continuous density $f_t\left(\cdot|\mathcal{F}_{t-1}\right)$ that satisfies $f_t(y|\mathcal{F}_{t-1}) \leq K < \infty$ and $|f_t(y'|\mathcal{F}_{t-1}) - f_t(y''|\mathcal{F}_{t-1})| \leq K\,|y' - y''|$, (ii) $\mathbb{E}\left[|Y_t|^{4+\delta}\right] \leq K < \infty$, for some $0 < \delta < 1$.

(C) There exists a neighborhood of $\boldsymbol{\theta}^0$, $\mathcal{N}\left(\boldsymbol{\theta}^0\right)$, such that for all t we have (i) $|1/e_t(\theta)| \leq K < \infty$, $\forall\ \theta \in \mathcal{N}\left(\boldsymbol{\theta}^0\right)$, (ii) there exist some (possibly stochastic) $\mathcal{F}_{t-1}$-measurable functions $V(\mathcal{F}_{t-1})$, $V_1(\mathcal{F}_{t-1})$, $H_1(\mathcal{F}_{t-1})$, $V_2(\mathcal{F}_{t-1})$, $H_2(\mathcal{F}_{t-1})$ that satisfy $\forall\ \boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}^0)$: $|v_t(\boldsymbol{\theta})| \leq V(\mathcal{F}_{t-1})$, $\|\nabla v_t(\boldsymbol{\theta})\| \leq V_1(\mathcal{F}_{t-1})$, $\|\nabla e_t(\boldsymbol{\theta})\| \leq H_1(\mathcal{F}_{t-1})$, $\left\|\nabla^2 v_t(\boldsymbol{\theta})\right\| \leq V_2(\mathcal{F}_{t-1})$, and $\left\|\nabla^2 e_t(\boldsymbol{\theta})\right\| \leq H_2(\mathcal{F}_{t-1})$.

(D) For some $0 < \delta < 1$ and for all t we have (i) $\mathbb{E}\left[V_1(\mathcal{F}_{t-1})^{3+\delta}\right]$, $\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{3+\delta}\right]$, $\mathbb{E}\left[V_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}}\right]$, $\mathbb{E}\left[H_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}}\right] \leq K$, (ii) $\mathbb{E}\left[V(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1})^{2+\delta}\right] \leq K$, (iii) $\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{1+\delta} H_2(\mathcal{F}_{t-1})\,|Y_t|^{2+\delta}\right]$, $\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{3+\delta}\,|Y_t|^{2+\delta}\right] \leq K$.

(E) The matrix $\mathbf{D}_0$ defined in Theorem 2 is (strictly) positive definite for T sufficiently large.

(F) $\left\{\left[Y_t, v_t\left(\boldsymbol{\theta}^0\right), e_t\left(\boldsymbol{\theta}^0\right), \nabla' v_t\left(\boldsymbol{\theta}^0\right), \nabla' e_t\left(\boldsymbol{\theta}^0\right)\right]\right\}$ is α-mixing with $\sum_{m=1}^{\infty} \alpha\,(m)^{(q-2)/q} < \infty$ for some $q > 2$.

(G) For any T, $\sup_{\boldsymbol{\theta}\in\Theta}\ \sum_{t=1}^{T} \mathbf{1}\left\{Y_t = v_t\left(\boldsymbol{\theta}\right)\right\} \leq K$ a.s.

Most of the above assumptions are standard. Assumption 2(A)(ii) imposes that the VaR is negative, but given our focus on the left-tail ($\alpha < 0.5$) of asset returns, this is not likely a binding constraint. Assumptions 2(B)–(E) are similar to those in Engle and Manganelli (2004a). Assumption 2(B)(ii) requires at least $4+\delta$ moments of returns to exist, however 2(D) may actually increase the number of required moments, depending on the VaR-ES model employed. Our requirement of at least $4 + \delta$ moments of returns allows returns to be fat tailed, but not without limit: it rules out applications where kurtosis is not defined, for example Student's t distributions with degrees of freedom of four or less. (In our simulation study below, we show that the theory here has good finite sample properties when using a Skew t with five degrees of freedom.) Assumptions 2(C)–(D)

are conditions on the magnitude of the VaR and ES paths, as well as first and second derivatives of these, making them somewhat hard to interpret. In the Supplemental Appendix we show that for a GARCH process these reduce to moment conditions on the observed returns. Assumption 2(F) is a standard condition on the amount of time series dependence, and allows us to invoke a CLT of Hall and Heyde (1980). Assumption 2(G) limits the number of exact equalities of realized returns and fitted VaR values; given assumption 2(B), in linear models $K = \dim(\boldsymbol{\theta})$, while in nonlinear models it may be that $K < \dim(\boldsymbol{\theta})$.

Theorem 2 (Asymptotic Normality) *Under Assumptions 1 and 2, we have*

$$\sqrt{T}\mathbf{A}_0^{-1/2}\mathbf{D}_0(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^0) \xrightarrow{d} N(0, I) \text{ as } T \to \infty \tag{30}$$

where

$$\mathbf{D}_0 = \mathbb{E}\left[\frac{f_t\left(v_t(\boldsymbol{\theta}^0)|\mathcal{F}_{t-1}\right)}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0) + \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla' e_t(\boldsymbol{\theta}^0)\nabla e_t(\boldsymbol{\theta}^0)\right] \tag{31}$$

$$\mathbf{A}_0 = \mathbb{E}\left[g_t(\boldsymbol{\theta}^0)g_t(\boldsymbol{\theta}^0)'\right] \tag{32}$$

$$\begin{aligned} g_t(\boldsymbol{\theta}) &= \frac{\partial L\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right); \alpha\right)}{\partial \boldsymbol{\theta}} \\ &= \nabla' v_t(\boldsymbol{\theta})\frac{1}{-e_t(\boldsymbol{\theta})}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t(\boldsymbol{\theta})\right\} - 1\right) \\ &+ \nabla' e_t(\boldsymbol{\theta})\frac{1}{e_t(\boldsymbol{\theta})^2}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t(\boldsymbol{\theta})\right\}\left(v_t(\boldsymbol{\theta}) - Y_t\right) - v_t(\boldsymbol{\theta}) + e_t(\boldsymbol{\theta})\right) \end{aligned} \tag{33}$$

An outline of the proof of this theorem is given in Appendix A, and the detailed lemmas underlying it are provided in the supplemental appendix. The proof of Theorem 2 builds on Huber (1967), Weiss (1991) and Engle and Manganelli (2004a), who focused on the estimation of quantiles.

Finally, we present a result for estimating the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_T$, thereby enabling the reporting of standard errors and confidence intervals.

Assumption 3 *(A) The deterministic positive sequence c_T satisfies $c_T = o(1)$ and $c_T^{-1} = o(T^{1/2})$.*

(B)(i) $T^{-1}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^0)g_t(\boldsymbol{\theta}^0)' - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$, where $\mathbf{A}_0$ is defined in Theorem 2.

(ii) $T^{-1}\sum_{t=1}^{T}\frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla' e_t(\boldsymbol{\theta}^0)\nabla e_t(\boldsymbol{\theta}^0) - \mathbb{E}\left[\frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla' e_t(\boldsymbol{\theta}^0)\nabla e_t(\boldsymbol{\theta}^0)\right] \xrightarrow{p} \mathbf{0}$.

(iii) $T^{-1}\sum_{t=1}^{T}\frac{f_t(v_t(\boldsymbol{\theta}^0)|\mathcal{F}_{t-1})}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0) - \mathbb{E}\left[\frac{f_t(v_t(\boldsymbol{\theta}^0)|\mathcal{F}_{t-1})}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0)\right] \xrightarrow{p} \mathbf{0}$.

Theorem 3 *Under Assumptions 1-3,* $\hat{\mathbf{A}}_T - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$ *and* $\hat{\mathbf{D}}_T - \mathbf{D}_0 \xrightarrow{p} \mathbf{0}$*, where*

$$\hat{\mathbf{A}}_T = T^{-1} \sum_{t=1}^{T} g_t(\hat{\boldsymbol{\theta}}_T) g_t(\hat{\boldsymbol{\theta}}_T)'$$

$$\hat{\mathbf{D}}_T = T^{-1} \sum_{t=1}^{T} \left\{ \frac{1}{2c_T} \mathbf{1}\left\{ \left| Y_t - v_t\left(\hat{\boldsymbol{\theta}}_T\right) \right| < c_T \right\} \frac{\nabla' v_t\left(\hat{\boldsymbol{\theta}}_T\right) \nabla v_t\left(\hat{\boldsymbol{\theta}}_T\right)}{-\alpha e_t\left(\hat{\boldsymbol{\theta}}_T\right)} + \frac{\nabla' e_t\left(\hat{\boldsymbol{\theta}}_T\right) \nabla e_t\left(\hat{\boldsymbol{\theta}}_T\right)}{e_t^2\left(\hat{\boldsymbol{\theta}}_T\right)} \right\}$$

This result extends Theorem 3 in Engle and Manganelli (2004a) from dynamic VaR models to dynamic joint models for VaR and ES. The key choice in estimating the asymptotic covariance matrix is the bandwidth parameter in Assumption 3(A). In our simulation study below we set this to $T^{-1/3}$ and we find that this leads to satisfactory finite-sample properties.

The results here extend some very recent work in the literature: Dimitriadis and Bayer (2017) consider VaR-ES regression, but focus on *iid* data and linear specifications. These authors also consider a variety of FZ loss functions, in contrast with our focus on the FZ0 loss function, and they consider both M and GMM estimation, while we focus only on M estimation. Barendse (2017) considers "interquantile expectation regression," which nests VaR-ES regression as a special case. He allows for time series data, but imposes that the models are linear. Our framework allows for time series data and nonlinear models.

4 Simulation study

In this section we investigate the finite-sample accuracy of the asymptotic theory for dynamic ES and VaR models presented in the previous section. For ease of comparison with existing studies of related models, such as volatility and VaR models, we consider a GARCH(1,1) for the DGP, and estimate the parameters by FZ loss minimization. Specifically, the DGP is

$$Y_t = \sigma_t \eta_t \tag{34}$$

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma Y_{t-1}^2$$

$$\eta_t \sim iid\ F_\eta(0,1) \tag{35}$$

We set the parameters of this DGP to $(\omega, \beta, \gamma) = (0.05, 0.9, 0.05)$. We consider two choices for the distribution of η_t: a standard Normal, and the standardized skew t distribution of Hansen (1994),

with degrees of freedom (ϑ) and skewness (λ) parameters in the latter set to $(5, -0.5)$. Under this DGP, the ES and VaR are proportional to σ_t, with

$$(\text{VaR}_t^\alpha, \text{ES}_t^\alpha) = (a_\alpha, b_\alpha)\, \sigma_t \tag{36}$$

We make the dependence of the coefficients of proportionality (a_α, b_α) on α explicit here, as we consider a variety of values of α in this simulation study: $\alpha \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$. Interest in VaR and ES from regulators focuses on the smaller of these values of α, but we also consider the larger values to better understand the properties of the asymptotic approximations at various points in the tail of the distribution.

For a standard Normal distribution, with CDF and PDF denoted Φ and ϕ, we have:

$$\begin{aligned} a_\alpha &= \Phi^{-1}(\alpha) \\ b_\alpha &= -\phi\left(\Phi^{-1}(\alpha)\right)/\alpha \end{aligned} \tag{37}$$

For Hansen's skew t distribution we can obtain VaR (a_α) from the inverse CDF, which is available in closed form. We obtain a closed-form expression for ES (b_α) by extending results in Dobrev, et al. (2017) which provides analyical expressions for ES for (symmetric) Student's t random variables. Details are presented in Appendix B.4. As noted in Section 2.5, FZ loss minimization does not allow us to identify ω in the GARCH model, and in our empirical work we set this parameter to one, however to facilitate comparisons of the accuracy of estimates of (a_α, b_α) in our simulation study we instead set ω at its true value. This is done without loss of generality and merely eases the presentation of the results. To match our empirical application, we replace the parameter a_α with $c_\alpha = a_\alpha/b_\alpha$, and so our parameter vector becomes $(\beta, \gamma, b_\alpha, c_\alpha)$.

We consider two sample sizes, $T \in \{2500, 5000\}$ corresponding to 10 and 20 years of daily returns respectively. These large sample sizes enable us to consider estimating models for quantiles as low as 1%, which are often used in risk management. We repeat all simulations 1000 times. To mitigate sensitivity to starting values, we initially estimate all models using a "smoothed" version of the FZ0 loss function, and use the resulting estimate as the starting value for the estimation problem using the original, "unsmoothed," FZ0 loss function. Details are in Appendix C.

Table 1 presents results for the estimation of this model on standard Normal innovations, and Table 2 presents corresponding results for skew t innovations. The top row of each panel present the true parameter values, with the latter two parameters changing across α. The second row presents the median estimated parameter across simulations, and the third row presents the average bias in the estimated parameter. Both of these measures indicate that the parameter estimates are nicely centered on the true parameter values. The penultimate row presents the cross-simulation standard deviations of the estimated parameters, and we observe that these decrease with the sample size and increase as we move further into the tails (i.e., as α decreases), both as expected. Comparing the standard deviations across Tables 1 and 2, we also note that they are higher for skew t innovations than Normal innovations, again as expected.

The last row in each panel presents the coverage probabilities for 95% confidence intervals for each parameter, constructed using the estimated standard errors, with bandwidth parameter $c_T = \lfloor T^{-1/3} \rfloor$. For $\alpha \geq 0.05$ we see that the coverage is reasonable, ranging from around 0.88 to 0.96. For $\alpha = 0.025$ or $\alpha = 0.01$ the coverage tends to be too low, particularly for the smaller sample size. Thus some caution is required when interpreting the standard errors for the models with the smallest values of α. In Table S1 of the Supplemental Appendix we present results for (Q)MLE for the GARCH model corresponding to the results in Tables 1 and 2, using the theory of Bollerslev and Wooldridge (1992), and in Tables S2 and S3 we present results for CAViaR estimation of this model, using the "tick" loss function and the theory of Engle and Manganelli (2004a).[7] We find that (Q)MLE has better finite sample properties than FZ minimization, but CAViaR estimation has slightly worse properties than FZ minimization.

Table 3 presents results for $T = 500$, which is relatively short given our interest in tail events, but may be of interest when only limited data are available or when structural breaks are suspected.

[7] In (Q)MLE, the parameters to be estimated are (ω, β, γ), and they are obtained by maximizing the sample average of the Normal log-likelihood. In "CAViaR" estimation, the parameters are $(\omega, \beta, \gamma, a_\alpha)$ and they are obtained by minimizing the sample average of the "tick" loss function, defined as $L(y, v; \alpha) = (\mathbf{1}\{y \leq v\} - \alpha)(v - y)$. Like FZ estimation, in the CAViaR approach we find that a_α and ω are not separately identified. As for the study of FZ estimation, we set ω to its true value to facilitate interpretation of the results, and estimate the remaining three parameters.

We see here that the estimator remains approximately unbiased, however inference (e.g., through confidence intervals) is less reliable with this short sample.

[INSERT TABLES 1–3 ABOUT HERE]

In Table 4 we compare the efficiency of FZ estimation relative to (Q)MLE and to CAViaR estimation, for the parameters that all three estimation methods have in common, namely (β, γ). As expected, when the innovations are standard Normal, FZ estimation is substantially less efficient than MLE, however when the innovations are skew t the loss in efficiency drops and for some values of α FZ estimation is actually more efficient than QMLE. This switch in the ranking of the competing estimators is qualitatively in line with results in Francq and Zakoïan (2015). In Panel B of Table 4, we see that FZ estimation is generally, though not uniformly, more efficient than CAViaR estimation.

[INSERT TABLE 4 ABOUT HERE]

In many applications, interest is focused on the forecasted values of VaR and ES rather than the estimated parameters of the models generating these forecasts. To study this, Table 5 presents results on the accuracy of the fitted VaR and ES estimates for the three estimation methods: (Q)MLE, CAViaR and FZ estimation. We consider the same two DGPs as above, and two others that represent more challenging environments for QMLE. In the two additional DGPs, we assume the same mean and volatility dynamics as before, and we additionally allow the degrees of freedom (ϑ) and skewness (λ) parameters in the skew t distribution to vary in such a way as to either "offset" or "amplify" the dynamics in volatility, resulting in VaR and ES series that are either approximately constant, or proportional to the conditional variance rather than the conditional standard deviation. These two simulation designs represent simple ways to obtain dynamics in VaR and ES that are "far" from the dynamics in volatility, and is an environment where QMLE would be expected to perform poorly. Details are provided in Appendix D.

To obtain estimates of VaR and ES from the (Q)ML estimates, we follow common empirical practice and compute the sample VaR and ES of the estimated standardized residuals. The columns

labeled *MAE* present the mean absolute error from (Q)MLE, and in the next two columns of each panel we present the *relative MAE* of CAViaR and FZ to (Q)MLE.

For Normal innovations, reported in Panel A, MLE is the most accurate estimation method, as expected. Averaging across values of α, CAViaR is about 40% worse, while FZ is about 30% worse. For skew t innovations, reported in Panel B, the gap in performance closes somewhat, with CAViaR and FZ performing about 24% and 16% worse than QMLE. In Panels C and D we consider challenging environments for QMLE, where the dynamics in volatility, which is the focus in QMLE, are very different from those in VaR and ES, which are the focus in FZ estimation. Unsurprisingly, QMLE does poorly in this case compared with FZ estimation, with MAE ratios (averaging across α) of 0.41 and 0.61 in these two panels, indicating that FZ does between 1.5 and 2.5 times better than QMLE in these simulation designs. CAViaR also outperforms QMLE in these designs, with average MAE ratios of 0.50 and 0.65.

Overall, these simulation results show that the asymptotic results of the previous section provide reasonable approximations in finite samples, with the approximations improving for larger sample sizes and less extreme values of α. Compared with MLE, estimation by FZ loss minimization is less accurate when the innovations are Normal or skew t, but when the dynamics in VaR and ES are different from those in volatility, the benefits of FZ estimation becomes apparent. Across all simulation designs, we find that FZ estimation is generally more accurate than estimation using the CAViaR approach of Engle and Manganelli (2004a), likely attributable to the fact that FZ estimation draws on information from two tail measures, VaR and ES, while CAViaR was designed to only model VaR.

[INSERT TABLE 5 ABOUT HERE]

5 Forecasting equity index ES and VaR

We now apply the models discussed in Section 2 to the forecasting of ES and VaR for daily returns on four international equity indices. We consider the S&P 500 index, the Dow Jones Industrial Average, the NIKKEI 225 index of Japanese stocks, and the FTSE 100 index of UK stocks. Our

sample period is 1 January 1990 to 31 December 2016, yielding between 6,630 and 6,805 observations per series (the exact numbers vary due to differences in holidays and market closures). In our out-of-sample analysis, we use the first ten years for estimation, and reserve the remaining 17 years for evaluation and model comparison.

Table 6 presents full-sample summary statistics on these four return series. Average annualized returns range from -2.7% for the NIKKEI to 7.2% for the DJIA, and annualized standard deviations range from 17.0% to 24.7%. All return series exhibit mild negative skewness (around -0.15) and substantial kurtosis (around 10). The lower two panels of Table 6 present the sample VaR and ES for four choices of α.

Table 7 presents results from standard time series models estimated on these return series over the in-sample period (Jan 1990 to Dec 1999). In the first panel we present the estimated parameters of the optimal ARMA(p, q) models, where the choice of (p, q) is made using the BIC. We note that for three of the four series the optimal model includes just a constant, consistent with the well-known lack of predictability of daily equity returns. The second panel presents the parameters of the GARCH(1,1) model for conditional variance, and the lower panel presents the estimated parameters the skew t distribution applied to the standardized residuals. All of these parameters are broadly in line with values obtained by other authors for these or similar series.

[INSERT TABLES 6 AND 7 ABOUT HERE]

5.1 In-sample estimation

We now present estimates of the parameters of the models presented in Section 2, along with standard errors computed using the theory from Section 3. In the interests of space, we only report the parameter estimates for the S&P 500 index for $\alpha = 0.05$. The two-factor GAS model based on the FZ0 loss function is presented in the left panel of Table 8. This model allows for separate dynamics in VaR and ES, and we present the parameters for each of these risk measures in separate columns. We impose that the $\mathbf{B}$ matrix is diagonal for parsimony. We observe that the persistence of these processes is high, with the estimated b parameters equal to 0.993 and 0.994, similar to the persistence found in GARCH models (e.g., see Table 7). The model-implied average values

of VaR and ES are -1.589 and -2.313, similar to the sample values of these measures reported in Table 6. We observe that the coefficients on λ_e for both VaR and ES are small in magnitude and far from being statistically significant. The coefficients on λ_v are larger and more significant (the t-statistics are -2.95 and -2.58). The overall imprecision from the coefficients on the four forcing variables suggests that this model is over-parameterized. For example, proportionality of v_t and e_t would suggest that a one-factor model is sufficient. We can formally test for this in the context of the two-factor model by testing that $w_e/w_v = a_{ev}/a_{vv} = a_{ee}/a_{ve} \cap b_v = b_e$. We obtain a p-value of 0.77 for this restriction, indicating no evidence against proportionality.

The right panel of Table 8 shows three one-factor models for ES and VaR. The first is the one-factor GAS model, which is nested in the two-factor model presented in the left panel. We see a slight loss in fit (the average loss is slightly greater) but the parameters of this model are estimated with greater precision. The one-factor GAS model fits better than the GARCH model estimated via FZ loss minimization (reported in the penultimate column).[8] The "hybrid" model, augmenting the one-factor GAS model with a GARCH-type forcing variable, fits better than the other one-factor models, and also slightly better than the larger two-factor GAS model, and we observe that the coefficient on the GARCH forcing variable (δ) is significantly different from zero (with a t-statistic of 9.55). The computation times for these models is reported in the bottom row of Table 8; for comparison, the computation time for QML estimation of the GARCH model is 0.39 seconds.

[INSERT TABLE 8 ABOUT HERE]

5.2 Out-of-sample forecasting

We now turn to the out-of-sample (OOS) forecast performance of the models discussed above, as well as some competitor models from the existing literature. We will focus initially on the results for

[8] Recall that in all of the one-factor models, the intercept (ω) in the GAS equation is unidentified. We fix it at zero for the GAS-1F and Hybrid models, and at one for the GARCH-FZ model. This has no impact on the fit of these models for VaR and ES, but it means that we cannot interpret the estimated (a, b) parameters as the VaR and ES of the standardized residuals, and we no longer expect the estimated values to match the sample estimates in Table 6.

$\alpha = 0.05$, given the focus on that percentile in the extant VaR literature. (Results for other values of α are considered below, with details provided in the supplemental appendix.) We will consider a total of ten models for forecasting ES and VaR. Firstly, we consider three rolling window methods, using window lengths of 125, 250 and 500 days. We next consider ARMA-GARCH models, with the ARMA model orders selected using the BIC, and assuming that the distribution of the innovations is standard Normal or skew t, or estimating it nonparametrically using the sample ES and VaR of the estimated standardized residuals. Finally we consider four new semiparametric dynamic models for ES and VaR: the two-factor GAS model presented in Section 2.2, the one-factor GAS model presented in Section 2.3, a GARCH model estimated using FZ loss minimization, and the "hybrid" GAS/GARCH model presented in Section 2.5. We estimate these models using the first ten years as our in-sample period, and retain those parameter estimates throughout the OOS period.

In Figure 4 below we plot the fitted 5% ES and VaR for the S&P 500 return series, using three models: the rolling window model using a window of 125 days, the GARCH-EDF model, and the one-factor GAS model. This figure covers both the in-sample and out-of-sample periods. The figure shows that the average ES was estimated at around -2%, rising as high as around -1% in the mid 90s and mid 00s, and falling to its most extreme values of around -10% during the financial crisis in late 2008. Thus, like volatility, ES fluctuates substantially over time.

Figure 5 zooms in on the last two years of our sample period, to better reveal the differences in the estimates from these models. We observe the usual step-like movements in the rolling window estimate of VaR and ES, as the more extreme observations enter and leave the estimation window. Comparing the GARCH and GAS estimates, we see how they differ in reacting to returns: the GARCH estimates are driven by lagged squared returns, and thus move stochastically each day. The GAS estimates, on the other hand, only use information from returns when the VaR is violated, and on other days the estimates revert deterministically to the long-run mean. This generates a smoother time series of VaR and ES estimates. We investigate below which of these estimates provides a better fit to the data.

[INSERT FIGURES 4 AND 5 ABOUT HERE]

The left panel of Table 9 presents the average OOS losses, using the FZ0 loss function from equation (6), for each of the ten models, for the four equity return series. The lowest values in each column are highlighted in bold, and the second-lowest are in italics. We observe that the one-factor GAS model, labelled FZ1F, is the preferred model for the two US equity indices, while the Hybrid model is the preferred model for the NIKKEI and FTSE indices. The worst model is the rolling window with a window length of 500 days.

While average losses are useful for an initial look at OOS forecast performance, they do not reveal whether the gains are statistically significant. Table 10 presents Diebold-Mariano t-statistics on the loss differences, for the S&P 500 index. Corresponding tables for the other three equity return series are presented in Table S4 of the supplemental appendix. The tests are conducted as "row model minus column model" and so a positive number indicates that the column model outperforms the row model. The column "FZ1F" corresponding to the one-factor GAS model contains all positive entries, revealing that this model out-performed all competing models. This outperformance is strongly significant for the comparisons to the rolling window forecasts, as well as the GARCH model with Normal innovations. The gains relative to the GARCH model with skew t or nonparametric innovations are not significant, with DM t-statistics of 1.79 and 1.53 respectively. Similar results are found for the best models for each of the other three equity return series. Thus the worst models are easily separated from the better models, but the best few models are generally not significantly different. The supplemental appendix presents results analogous to Table 9, but with alpha=0.025, which is the value for ES that is the focus of the Basel III accord. The rankings and results are qualitatively similar to those for alpha=0.05 discussed here.

[INSERT TABLES 9 AND 10 ABOUT HERE]

To complement the study of the relative performance of these models for ES and VaR, we now consider goodness-of-fit tests for the OOS forecasts of VaR and ES. Under correct specification of the model for VaR and ES, we know that

$$\mathbb{E}_{t-1}\left[\begin{array}{c} \partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)/\partial v_t \\ \partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)/\partial e_t \end{array}\right] = 0 \tag{38}$$

and we note that this implies that $\mathbb{E}_{t-1}[\lambda_{v,t}] = \mathbb{E}_{t-1}[\lambda_{e,t}] = 0$, where $(\lambda_{v,t}, \lambda_{e,t})$ are defined in equations (11)-(12). Thus the variables $\lambda_{v,t}$ and $\lambda_{e,t}$ can be considered as a form of "generalized residual" for this model. To mitigate the impact of serial correlation in these measures (which comes through the persistence of v_t and e_t) we use standardized versions of these residuals:

$$\begin{aligned} \lambda^s_{v,t} &\equiv \frac{\lambda_{v,t}}{v_t} = \mathbf{1}\{Y_t \leq v_t\} - \alpha \\ \lambda^s_{e,t} &\equiv \frac{\lambda_{e,t}}{e_t} = \frac{1}{\alpha}\mathbf{1}\{Y_t \leq v_t\}\frac{Y_t}{e_t} - 1 \end{aligned} \tag{39}$$

These standardized generalized residuals are also conditionally mean zero under correct specification, and we note that the standardized residual for VaR is simply the demeaned "hit" variable, which is the focus of well-known tests from the VaR literature, see Christoffersen (1998) and Engle and Manganelli (2004a). We adopt the "dynamic quantile (DQ)" testing approach of Engle and Manganelli (2004a), which is based on simple regressions of these generalized residuals on elements of the information set available at the time the forecast was made. Consider, then the following "DQ" and "DES" regressions:

$$\begin{aligned} \lambda^s_{v,t} &= a_0 + a_1\lambda^s_{v,t-1} + a_2 v_t + u_{v,t} \\ \lambda^s_{e,t} &= b_0 + b_1\lambda^s_{e,t-1} + b_2 e_t + u_{e,t} \end{aligned} \tag{40}$$

where $\mathbf{a} = [a_0, a_1, a_2]'$ and $\mathbf{b} = [b_0, b_1, b_2]'$ are the parameters of the regression and $u_{v,t}$ and $u_{e,t}$ are the regression residuals. We test forecast optimality by testing that all parameters in these regressions are zero, against the usual two-sided alternative. Similar "conditional calibration" tests are presented in Nolde and Ziegel (2017). One could also consider a joint test of both of the above null hypotheses, however we will focus on these separately so that we can determine which variable is well/poorly specified.

The right two panels of Table 9 present the p-values from the tests of the goodness-of-fit of the VaR and ES forecasts. Entries greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and entries between 0.05 and 0.10 are in italics. For the S&P 500 index and the DJIA, we see that only one model passes the ES tests: the two-factor GAS model, while no model passes the VaR tests. For the NIKKEI we see that all of the dynamic models pass these two

tests, while all three of the rolling window models fail. For the FTSE index, on the other hand, we see that all ten models considered here fail both the goodness-of-fit tests. The outcomes for the NIKKEI and the FTSE each, in different ways, present good examples of the problem highlighted in Nolde and Ziegel (2017), that many different models may pass a goodness-of-fit test, or all models may fail, which makes discussing their relative performance difficult. To do so, one can look at Diebold-Mariano tests of differences in average loss, as we do in Table 10.

Finally, in Table 11 we look at the performance of these models across four values of α, to see whether the best-performing models change with how deep in the tails we are. We find that this is indeed the case: for $\alpha = 0.01$, the best-performing model across the four return series is the GARCH model estimated by FZ loss minimization, followed by the GARCH model with nonparametric residuals. These rankings also hold for $\alpha = 0.025$. For $\alpha = 0.05$ the two best models are the GARCH model with nonparametric residuals and the Hybrid model, while for $\alpha = 0.10$ the two best models are the Hybrid model and the one-factor GAS model. These rankings are perhaps related to the fact that the forcing variable in the GAS model depends on observing a violation of the VaR, and for very small values of α these violations occur only infrequently. In contrast, the GARCH model uses the information from the squared residual, and so information from the data moves the risk measures whether a VaR violation was observed or not. When α is not too small, the forcing variable suggested by the GAS model applied to the FZ loss function starts to out-perform.

[INSERT TABLE 11 ABOUT HERE]

6 Conclusion

With the implementation of the Third Basel Accord in the next few years, risk managers and regulators will place greater focus on expected shortfall (ES) as a measure of risk, complementing and partly substituting previous emphasis on Value-at-Risk (VaR). We draw on recent results from statistical decision theory (Fissler and Ziegel, 2016) to propose new dynamic models for ES and VaR. The models proposed are semiparametric, in that they impose parametric structures for the

dynamics of ES and VaR, but are agnostic about the conditional distribution of returns. We also present asymptotic distribution theory for the estimation of these models, and we verify that the theory provides a good approximation in finite samples. We apply the new models and methods to daily returns on four international equity indices, over the period 1990 to 2016, and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models.

The asymptotic theory presented in this paper facilitates considering a large number of extensions of the models presented here. Our models all focus on a single value for the tail probability (α), and extending these to consider multiple values simultaneously could prove fruitful. For example, one could consider the values 0.01, 0.025 and 0.05, to capture various points in the left tail, or one could consider 0.05 and 0.95 to capture both the left and right tails simultaneously. Another natural extension is to make use of exogenous information in the model; the models proposed here are all univariate, and one might expect that information from options markets, high frequency data, or news announcements to also help predict VaR and ES. We leave these interesting extensions to future research.

Appendix A: Proofs

Proof of Proposition 1. Theorem C.3 of Nolde and Ziegel (2017) shows that under the assumption that ES is strictly negative, the loss differences generated by a FZ loss function are homogeneous of degree zero iff $G_1(x) = \varphi_1 \mathbf{1}\{x \geq 0\}$ and $G_2(x) = -\varphi_2/x$ with $\varphi_1 \geq 0$ and $\varphi_2 > 0$. Denote the resulting loss function as $L^*_{FZ0}(Y, v, e; \alpha, \varphi_1, \varphi_2)$, and notice that:

$$\begin{aligned}
L^*_{FZ0}(Y, v, e; \alpha, \varphi_1, \varphi_2) &= \varphi_1 (\mathbf{1}\{Y \leq v\} - \alpha)(\mathbf{1}\{v \geq 0\} - \mathbf{1}\{Y \geq 0\}) \\
&\quad + \varphi_2 \left\{ -(\mathbf{1}\{Y \leq v\} - \alpha) \frac{1}{\alpha} \frac{v}{e} + \frac{1}{e}\left(\frac{1}{\alpha}\mathbf{1}\{Y \leq v\} Y - e\right) + \log(-e) \right\} \\
&= \varphi_1 (\mathbf{1}\{Y \leq v\} - \alpha)(\mathbf{1}\{v \geq 0\} - \mathbf{1}\{Y \geq 0\}) + \varphi_2 L_{FZ0}(Y, v, e; \alpha) \\
&= \varphi_2 L_{FZ0}(Y, v, e; \alpha) + \varphi_1 \alpha \mathbf{1}\{Y \geq 0\} \\
&\quad + \varphi_1 \mathbf{1}\{v \geq 0\}(\mathbf{1}\{Y \leq v\} - \alpha - \mathbf{1}\{0 \leq Y \leq v\})
\end{aligned}$$

Under the assumption that $v < 0$, the third term vanishes. The second term is purely a function of Y and so can be disregarded; we can set $\varphi_1 = 0$ without loss of generality. The first term is affected by a scaling parameter $\varphi_2 > 0$, and we can set $\varphi_2 = 1$ without loss of generality. Thus we obtain the L_{FZ0} given in equation (6). If v can be positive, then setting $\varphi_1 = 0$ is interpretable as fixing this shape parameter value at a particular value. ■

Proof of Theorem 1. The proof is based on Theorem 2.1 of Newey and McFadden (1994). We only need to show that $E[L_T(\cdot)]$ is uniquely minimized at $\boldsymbol{\theta}^0$, because the other assumptions of Newey and McFadden's theorem are clearly satisfied. By Corollary (5.5) of Fissler and Ziegel (2016), given Assumption 1(B)(iii) and the fact that our choice of the objective function L_{FZ0} satisfies the condition as in Corollary (5.5) of Fissler and Ziegel (2016), we know that $\mathbb{E}[L(Y_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}); \alpha) | \mathcal{F}_{t-1}]$ is uniquely minimized at $(\mathrm{VaR}_\alpha(Y_t|\mathcal{F}_{t-1}), \mathrm{ES}_\alpha(Y_t|\mathcal{F}_{t-1}))$, which equals $\big(v_t(\boldsymbol{\theta}^0), e_t(\boldsymbol{\theta}^0)\big)$ under correct specification. Combining this assumption and Assumption 1(B)(iv), we know that $\boldsymbol{\theta}^0$ is a unique minimizer of $\mathbb{E}[L_T(\cdot)]$, completing the proof. ■

Outline of proof of Theorem 2. We consider the population function $\lambda(\boldsymbol{\theta}) = \mathbb{E}[g_t(\boldsymbol{\theta})]$,

and take a mean-value expansion of $\lambda(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^0$. We show in Lemma 1 that:

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0) &= -\Lambda^{-1}(\boldsymbol{\theta}^0)\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^0)+o_p(1)\\ \text{where}\quad \Lambda(\boldsymbol{\theta}^*) &= \left.\frac{\partial \mathbb{E}\left[g_t(\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\end{aligned}$$

In the supplemental appendix we prove Lemma 1 by building on and extending Weiss (1991), who extends Huber (1967) to non-*iid* data. We draw on Weiss' Lemma A.1, and we verify that all five assumptions (N1-N5 in his notation) for that lemma are satisfied: N1, N2 and N5 are obviously satisfied given our Assumptions 1-2, and we show in Lemmas 3 - 6 that assumptions N3 and N4 are satistfied. Lemma 7 shows that a CLT applies for the sequence $\left\{T^{-1/2}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^0)\right\}$, with asymptotic covariance matrix $\mathbf{A}_0 = \mathbb{E}\left[g_t(\boldsymbol{\theta}^0)g_t(\boldsymbol{\theta}^0)'\right]$. We denote $\Lambda(\boldsymbol{\theta}^0)$ as $\mathbf{D}_0$, leading to the stated result. ■

Proof of Theorem 3. Given Assumption 3B(i) and the result in Theorem 1, the proof that $\hat{\mathbf{A}}_T - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$ is standard and omitted. Next, define

$$\tilde{\mathbf{D}}_T = T^{-1}\sum_{t=1}^{T}\{(2c_T)^{-1}\mathbf{1}\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\}\frac{1}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla v_t(\boldsymbol{\theta}^0)'\nabla v_t(\boldsymbol{\theta}^0) + \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^0)\}$$

To prove the result we will show that $\hat{\mathbf{D}}_T - \tilde{\mathbf{D}}_T = o_p(1)$ and $\tilde{\mathbf{D}}_T - \mathbf{D}_0 = o_p(1)$. Firstly, consider

$$\begin{aligned}\|\hat{\mathbf{D}}_T - \tilde{\mathbf{D}}_T\| \leq \big\|&(2Tc_T)^{-1}\\ &\times\sum_{t=1}^{T}\{(\mathbf{1}\{|Y_t - v_t(\hat{\boldsymbol{\theta}}_T)| < c_T\} - \mathbf{1}\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\})\frac{1}{-e_t(\hat{\boldsymbol{\theta}}_T)\alpha}\nabla v_t(\hat{\boldsymbol{\theta}}_T)'\nabla v_t(\hat{\boldsymbol{\theta}}_T)\\ &+\mathbf{1}\left\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\right\}\frac{1}{-e_t(\hat{\boldsymbol{\theta}}_T)\alpha}\left(\nabla v_t(\hat{\boldsymbol{\theta}}_T) - \nabla v_t(\boldsymbol{\theta}^0)\right)'\nabla v_t(\hat{\boldsymbol{\theta}}_T)\\ &+\mathbf{1}\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\}\left(\frac{1}{-\alpha e_t(\hat{\boldsymbol{\theta}}_T)} - \frac{1}{-\alpha e_t(\boldsymbol{\theta}^0)}\right)\nabla v_t(\boldsymbol{\theta}^0)'\nabla v_t(\hat{\boldsymbol{\theta}}_T)\\ &+\mathbf{1}\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\}\frac{1}{-\alpha e_t(\boldsymbol{\theta}^0)}\nabla v_t(\boldsymbol{\theta}^0)'(\nabla v_t(\hat{\boldsymbol{\theta}}_T) - \nabla v_t(\boldsymbol{\theta}^0))\Big\|\\ &+T^{-1}\sum_{t=1}^{T}\left\|\frac{1}{e_t(\hat{\boldsymbol{\theta}}_T)^2}\nabla e_t(\hat{\boldsymbol{\theta}}_T)'\nabla e_t(\hat{\boldsymbol{\theta}}_T) - \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^0)\right\|\end{aligned}$$

The last line above was shown to be $o_p(1)$ in the proof of Theorem 2. The difficult quantity in the first term (over the first six lines above) is the indicator, and following the same steps as in Engle

and Manganelli (2004a), that term is also $o_p(1)$. Next, consider $\tilde{\mathbf{D}}_T - \mathbf{D}_0$:

$$\begin{aligned}
\tilde{\mathbf{D}}_T - \mathbf{D}_0 &= \frac{1}{2Tc_T}\sum_{t=1}^{T}\left(\mathbf{1}\left\{\left|Y_t - v_t(\boldsymbol{\theta}^0)\right| < c_T\right\} - \mathbb{E}\left[\mathbf{1}\left\{\left|Y_t - v_t\left(\boldsymbol{\theta}^0\right)\right| < c_T\right\}|\mathcal{F}_{t-1}\right]\right) \\
&\times \frac{\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0)}{-e_t(\boldsymbol{\theta}^0)\alpha} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2c_T}\mathbb{E}[\mathbf{1}\{|Y_t - v_t(\boldsymbol{\theta}^0)| < c_T\}|\mathcal{F}_{t-1}]\frac{1}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0)\right. \\
&\left. - \mathbb{E}\left[\frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)\alpha}\nabla' v_t(\boldsymbol{\theta}^0)\nabla v_t(\boldsymbol{\theta}^0)\right]\right\} + o_p(1)
\end{aligned}$$

Following Engle and Manganelli (2004a), assumptions 1-3 are sufficient to show $\tilde{\mathbf{D}}_T - \mathbf{D}_0 = o_p(1)$ and the result follows. ■

Appendix B: Derivations

Appendix B.1: Generic calculations for the FZ0 loss function

The FZ0 loss function is:

$$L_{FZ0}(Y, v, e; \alpha) = -\frac{1}{\alpha e}\mathbf{1}\{Y \leq v\}(v - Y) + \frac{v}{e} + \log(-e) - 1 \tag{41}$$

Note that this is *not* homogeneous, as for any $k > 0$, $L_{FZ0}(kY, kv, ke; \alpha) = L_{FZ0}(Y, v, e; \alpha) + \log(k)$, but this loss function generates loss *differences* that are homogenous of degree zero, as the additive additional term above drops out.

We will frequently use the first derivatives of this loss function, and the second derivatives of the expected loss for an absolutely continuous random variable with density f and CDF F. These are (for $v \neq Y$):

$$\nabla_v \equiv \frac{\partial L_{FZ0}(Y, v, e; \alpha)}{\partial v} = -\frac{1}{\alpha e}(\mathbf{1}\{Y \leq v\} - \alpha) \equiv \frac{1}{\alpha v e}\lambda_v \tag{42}$$

$$\begin{aligned}
\nabla_e &\equiv \frac{\partial L_{FZ0}(Y, v, e; \alpha)}{\partial e} \\
&= \frac{1}{\alpha e^2}\mathbf{1}\{Y \leq v\}(v - Y) - \frac{v}{e^2} + \frac{1}{e} \\
&= \frac{v}{\alpha e^2}(\mathbf{1}\{Y \leq v\} - \alpha) - \frac{1}{e^2}\left(\frac{1}{\alpha}\mathbf{1}\{Y \leq v\}Y - e\right) \\
&\equiv \frac{-1}{\alpha e^2}(\lambda_v + \alpha\lambda_e)
\end{aligned} \tag{43}$$

where

$$\lambda_v \equiv -v\left(\mathbf{1}\left\{Y \leq v\right\} - \alpha\right) \tag{44}$$

$$\lambda_e \equiv \frac{1}{\alpha}\mathbf{1}\left\{Y \leq v\right\} Y - e \tag{45}$$

and

$$\begin{aligned}
\frac{\partial^2 \mathbb{E}\left[L_{FZ0}\left(Y, v, e; \alpha\right)\right]}{\partial v^2} &= -\frac{1}{\alpha e} f\left(v\right) && (46)\\
\frac{\partial^2 \mathbb{E}\left[L_{FZ0}\left(Y, v, e; \alpha\right)\right]}{\partial v \partial e} &= \frac{1}{\alpha e^2}\left(F\left(v\right) - \alpha\right) && (47)\\
&= 0, \text{ at the true value of } (v, e) \\
\frac{\partial^2 \mathbb{E}\left[L_{FZ0}\left(Y, v, e; \alpha\right)\right]}{\partial e^2} &= \frac{1}{e^2} - \frac{2}{\alpha e^3}\left\{\left(F\left(v\right) - \alpha\right) v - \left(\mathbb{E}\left[\mathbf{1}\left\{Y \leq v\right\} Y\right] - \alpha e\right)\right\} && (48)\\
&= \frac{1}{e^2}, \text{ at the true value of } (v, e)
\end{aligned}$$

Appendix B.2: Derivations for the one-factor GAS model for ES and VaR

Here we present the calculations to compute s_t and I_t for this model. Below we use:

$$\frac{\partial v}{\partial \kappa} = \frac{\partial^2 v}{\partial \kappa^2} = a \exp\left\{\kappa\right\} = v \tag{49}$$

$$\frac{\partial e}{\partial \kappa} = \frac{\partial^2 e}{\partial \kappa^2} = b \exp\left\{\kappa\right\} = e \tag{50}$$

And so we find (for $v_t \neq Y_t$)

$$\begin{aligned}
s_t &\equiv \frac{\partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)}{\partial \kappa_t} && (51)\\
&= \frac{\partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)}{\partial v_t}\frac{\partial v_t}{\partial \kappa_t} + \frac{\partial L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)}{\partial e_t}\frac{\partial e_t}{\partial \kappa_t} \\
&= \left\{-\frac{1}{\alpha e_t}\left(\mathbf{1}\left\{Y_t \leq v_t\right\} - \alpha\right)\right\} v_t \\
&\quad + \left\{-\frac{1}{e_t^2}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t\right\} Y_t - e_t\right) + \frac{v_t}{e_t^2}\frac{1}{\alpha}\left(\mathbf{1}\left\{Y_t \leq v_t\right\} - \alpha\right)\right\} e_t \\
&= -\frac{1}{e_t}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t\right\} Y_t - e_t\right) && (52)\\
&\equiv -\lambda_{et}/e_t && (53)
\end{aligned}$$

Thus, the λ_{vt} term drops out of s_t and we are left with $-\lambda_{et}/e_t$.

Next we calculate I_t :

$$\begin{aligned} I_t &\equiv \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial \kappa_t^2} \qquad (54) \\ &= \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t^2}\left(\frac{\partial v_t}{\partial \kappa_t}\right)^2 + \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t \partial e_t}\frac{\partial v_t}{\partial \kappa_t} \\ &\quad + \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial e_t^2}\left(\frac{\partial e_t}{\partial \kappa_t}\right)^2 + \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t \partial e_t}\frac{\partial e_t}{\partial \kappa_t} \\ &\quad + \frac{\partial \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t}\frac{\partial^2 v_t}{\partial \kappa_t^2} + \frac{\partial \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial e_t}\frac{\partial^2 e_t}{\partial \kappa_t^2} \end{aligned}$$

But note that under correct specification,

$$\frac{\partial^2 \mathbb{E}_{t-1}\left[L\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t \partial e_t} = \frac{\partial \mathbb{E}_{t-1}\left[L\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t} = \frac{\partial \mathbb{E}_{t-1}\left[L\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial e_t} = 0 \qquad (55)$$

and so the Hessian simplifies to:

$$\begin{aligned} I_t &= \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial v_t^2}\left(\frac{\partial v_t}{\partial \kappa_t}\right)^2 + \frac{\partial^2 \mathbb{E}_{t-1}\left[L_{FZ0}\left(Y_t, v_t, e_t; \alpha\right)\right]}{\partial e_t^2}\left(\frac{\partial e_t}{\partial \kappa_t}\right)^2 \qquad (56) \\ &= -\frac{1}{\alpha e_t} f_t\left(v_t\right) v_t^2 + 1 \qquad (57) \\ &= 1 - \frac{k_\alpha}{\alpha}\frac{a_\alpha}{b_\alpha}, \text{ since } f_t\left(v_t\right) = \frac{k_\alpha}{v_t} \text{ and } \frac{v_t}{e_t} = \frac{a_\alpha}{b_\alpha}, \text{ for this DGP.} \qquad (58) \end{aligned}$$

Thus although the Hessian could vary with time, as it is a derivative of the conditional expected loss, in this specification it simplifies to a (positive) constant.

Appendix B.3: ES and VaR in location-scale models

Dynamic location-scale models are widely used for asset returns and in this section we consider what such a specification implies for the dynamics of ES and VaR. Consider the following:

$$Y_t = \mu_t + \sigma_t \eta_t, \quad \eta_t \sim iid\ F_\eta\left(0, 1\right) \qquad (59)$$

where, for example, μ_t is some ARMA model and σ_t^2 is some GARCH model. For asset returns that follow equation (59) we have:

$$\begin{aligned} v_t &= \mu_t + a\sigma_t, \quad \text{where} \quad a = F_\eta^{-1}\left(\alpha\right) \qquad (60) \\ e_t &= \mu_t + b\sigma_t, \quad \text{where} \quad b = \mathbb{E}\left[\eta_t | \eta_t \leq a\right] \end{aligned}$$

and we we can recover (μ_t, σ_t) from (v_t, e_t):

$$\begin{bmatrix} \mu_t \\ \sigma_t \end{bmatrix} = \frac{1}{b-a} \begin{bmatrix} b & -a \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_t \\ e_t \end{bmatrix} \tag{61}$$

Thus under the conditional location-scale assumption, we can back out the conditional mean and variance from the VaR and ES. Next note that if $\mu_t = 0 \; \forall \; t$, then $v_t = c \cdot e_t$, where $c = a/b \in (0, 1)$. Daily asset returns often have means that are close to zero, and so this restriction is one that may be plausible in the data. A related, though less plausible, restriction is that $\sigma_t = \bar{\sigma} \; \forall \; t$, and in that case we have the simplification that $v_t = d + e_t$, where $d = (a - b)\, \bar{\sigma} > 0$.

Appendix B.4: VaR and ES for Hansen's skew t random variables

The VaR for Hansen's (1994) skew t variable is can be obtained using an expression for the inverse CDF of the skew t distribution presented in Jondeau and Rockinger (2003). In a recent paper, Dobrev et al. (2017) present an analyical expression for the expected shortfall for a Student's t random variable, X, with degrees of freedom $\nu > 1$:

$$ES_x(\alpha; \nu) = \frac{\nu^{\nu/2}}{2\alpha\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\left(VaR_x(\alpha; \nu)\right)^2 + \nu \right)^{-(\nu-1)/2} \tag{62}$$

where $VaR_x(\alpha; \nu) = F_x^{-1}(\alpha; \nu)$ is the α-quantile of a Student's $t(v)$ variable. The VaR for a standardized (unit variance) Student's t variable, Y, is simply:

$$ES_y(\alpha; \nu) = \sqrt{\frac{\nu - 2}{\nu}} ES_x(\alpha; \nu) \tag{63}$$

We use the connection between the PDF of a standardized Student's t random variable and Hansen's (1994) skew t variable, Z, to obtain an analogous expression for the expected shortfall of a skew t random variable. As in Hansen (1994), define

$$a \equiv 4\lambda c \left(\frac{\nu - 2}{\nu - 1} \right), \;\; b \equiv \sqrt{1 + 3\lambda^2 - a^2}, \;\; c \equiv \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\,(\nu - 2)}} \tag{64}$$

Define

$$\begin{aligned} ES_z^*(\alpha; \nu, \lambda) &= \frac{\tilde{\alpha}}{\alpha}(1 - \lambda) \left(-\frac{a}{b} + \frac{1 - \lambda}{b} ES_y(\alpha; \nu) \right) \\ \text{where} \quad \tilde{\alpha} &\equiv F_z\left(\frac{b}{1 - \lambda} \left(VaR_y(\alpha; \nu) + \frac{a}{b} \right); \nu, 0 \right) \end{aligned} \tag{65}$$

where $F_z(\cdot;\nu,\lambda)$ is the CDF of the skew t distribution with parameters (ν,λ). Let $VaR_z(\alpha;\nu,\lambda) = F_z^{-1}(\alpha;\nu,\lambda)$. It can then be shown that

$$ES_z(\alpha;\nu,\lambda) = \begin{cases} ES_z^*(\alpha;\nu,\lambda), & VaR_z(\alpha;\nu,\lambda) \leq -a/b \\ \frac{1-\alpha}{\alpha} ES_z^*(1-\alpha;\nu,-\lambda), & VaR_z(\alpha;\nu,\lambda) > -a/b \end{cases} \tag{66}$$

Note that when $\lambda = 0$ this simplifies to a symmetric unit-variance Student's t variable, and we recover the expression above, i.e., $ES_z(\alpha;\nu,0) = ES_y(\alpha;\nu)$. A Matlab function for the VaR and ES of a skew t variable is available at the link given in the first footnote of this paper.

Appendix C: Estimation using the FZ0 loss function

The FZ0 loss function, equation (6), involves the indicator function $\mathbf{1}\{Y_t \leq v_t\}$ and so necessitates the use of a numerical search algorithm that does not rely on differentiability of the objective function; we use the function `fminsearch` in Matlab. However, in preliminary simulation analyses we found that this algorithm was sensitive to the starting values used in the search. To overcome this, we initially consider a "smoothed" version of the FZ0 loss function, where we replace the indicator variable with a Logistic function:

$$\tilde{L}_{FZ0}(Y,v,e;\alpha,\tau) = -\frac{1}{\alpha e}\Gamma(Y_t,v_t;\tau)(v-Y) + \frac{v}{e} + \log(-e) - 1 \tag{67}$$

$$\text{where} \quad \Gamma(Y_t,v_t;\tau) \equiv \frac{1}{1+\exp\{\tau(Y_t-v_t)\}}, \text{ for } \tau > 0 \tag{68}$$

where τ is the smoothing parameter, and the smoothing function Γ converges to the indicator function as $\tau \to \infty$. In GAS models that involve an indicator function in the forcing variable, we alter the forcing variable in the same way, to ensure that the objective function as a function of $\boldsymbol{\theta}$ is differentiable. In these cases the loss function *and* the model itself are slightly altered through this smoothing.

In our empirical implementation, we obtain "smart" (or "warm") starting values by first estimating the model using the "smoothed FZ0" loss function with $\tau = 5$. This choice of τ gives some smoothing for values of Y_t that are roughly within ± 1 of v_t. Call the resulting parameter estimate $\tilde{\boldsymbol{\theta}}_T^{(5)}$. Since this objective function is differentiable, we can use more familiar gradient-based numerical search algorithms, such as `fminunc` or `fmincon` in Matlab, which are often less sensitive to

starting values. We then re-estimate the model, using $\tilde{\boldsymbol{\theta}}_T^{(5)}$ as the starting value, setting $\tau = 20$ and obtain $\tilde{\boldsymbol{\theta}}_T^{(20)}$. This value of τ smoothes values of Y_t within roughly ± 0.25 of v_t, and so this objective function is closer to the true objective function. Finally, we use $\tilde{\boldsymbol{\theta}}_T^{(20)}$ as the starting value in the optimization of the actual FZ0 objective function, with no artificial smoothing, using the function `fminsearch`, and obtain $\hat{\boldsymbol{\theta}}_T$. We found that this approach largely eliminated the sensitivity to starting values.

Appendix D: Dynamics in the Skew t distribution

For each parameter vector (ϑ, λ) of the Skew t distribution, we define $h(\vartheta, \lambda) \equiv [h_v(\vartheta, \lambda), h_e(\vartheta, \lambda)] = (v, e)$ to be the VaR and ES at a given value of α.Given the functional form of the Skew t distribution, not all pairs of VaR and ES are attainable from a set of parameters (ϑ, λ) and so the function h is not invertible everywhere. We define a pseudo-inverse of this mapping as

$$h^{(-1)}(v, e) \equiv \arg\min_{(\vartheta,\lambda)} \; (h_v(\vartheta, \lambda) - v)^2 + (h_e(\vartheta, \lambda) - e)^2 \tag{69}$$

In words, the pseudo-inverse returns the Skew t parameters (ϑ, λ) that lead to VaR and ES that are as close as possible, in a squared-error distance metric, to the target values (v, e).

To obtain dynamics in (ϑ, λ) that "offset" those in the GARCH volatility process, we set $(\vartheta_t, \lambda_t) = h^{(-1)}(\bar{v}/\sigma_t, \bar{e}/\sigma_t)$, and we fix $(\bar{v}, \bar{e}) = \left(\Phi^{-1}(\alpha), -\phi\left(\Phi^{-1}(\alpha)\right)/\alpha\right)$. If the pseudo-inverse was actually the proper inverse, we would find:

$$(v_t, e_t) = \sigma_t h(\vartheta_t, \lambda_t) = \sigma_t h\left(h^{(-1)}(\bar{v}/\sigma_t, \bar{e}/\sigma_t)\right) = (\bar{v}, \bar{e}) \;\; \forall \; t \tag{70}$$

In our simulation study, the time series of (v_t, e_t) in the "offsetting" case were approximately but not perfectly flat, e.g., for $\alpha = 0.05$, the min-max spreads for v_t and e_t were around 0.05, when $(\bar{v}, \bar{e}) = (-1.65, -2.06)$. As the dynamics in volatility are not completely offset, this is helpful for QMLE and the ratios reported in Panel C of Table 5 are better than if the dynamics could be offset completely.

To obtain "amplifying" dynamics, we set $(\vartheta_t, \lambda_t) = h^{(-1)}\left(\bar{a}\sigma_t, \bar{b}\sigma_t\right)$, and we fix $\left(\bar{a}, \bar{b}\right) = (\bar{v}, \bar{e})/(2\bar{\sigma})$. If $h^{(-1)}$ were a proper inverse this would lead to:

$$(v_t, e_t) = \sigma_t h(\vartheta_t, \lambda_t) = \sigma_t h\left(h^{(-1)}\left(\bar{a}\sigma_t, \bar{b}\sigma_t\right)\right) = \left(\bar{a}, \bar{b}\right)\sigma_t^2 = \left(\frac{\bar{v}\sigma_t}{2\bar{\sigma}}, \frac{\bar{e}\sigma_t}{2\bar{\sigma}}\right)\sigma_t \equiv (a_t, b_t)\,\sigma_t \tag{71}$$

and so the coefficients linking VaR and ES to conditional standard deviation are also increasing in the standard deviation, yielding the "amplifying" feature of this model. In our simulation study, the time series of (v_t, e_t) were almost perfectly linear in σ_t^2, with R^2 values from regressions of v_t and e_t on σ_t^2 of over 0.998 in all cases.

References

[1] Andersen, T.G., Bollerslev, T., Christoffersen, P., Diebold, F.X., 2006. Volatility and Correlation Forecasting, in (ed.s) G. Elliott, C.W.J. Granger, and A. Timmermann, *Handbook of Economic Forecasting*, Vol. 1. Elsevier, Oxford.

[2] Andrews, D.W.K., 1987, Consistency in nonlinear econometric models: ageneric uniform law of large numbers, *Econometrica*, 55, 1465–1471.

[3] Artzner, P., F. Delbaen, J.M. Eber and D. Heath, 1999, Coherent measures of risk, *Mathematical Finance*, 9, 203-228.

[4] Barendse, S., 2017, Interquantile Expectation Regression, Tinbergen Institute Discussion Paper, TI 2017-034/III.

[5] Basel Committee on Banking Supervision, 2010, Basel III: A Global Regulatory Framework for More Resiliant Banks and Banking Systems, Bank for International Settlements. `http://www.bis.org/publ/bcbs189.pdf`

[6] Bollerslev, T., 1986, Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, 31, 307-327.

[7] Bollerslev, T. and J.M. Wooldridge, 1992, Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances, *Econometric Reviews*, 11(2), 143-172.

[8] Cai, Z. and X. Wang, 2008, Nonparametric estimation of conditional VaR and expected shortfall, *Journal of Econometrics*, 147, 120-130.

[9] Creal, D.D., S.J. Koopman, and A. Lucas, 2013, Generalized Autoregressive Score Models with Applications, *Journal of Applied Econometrics*, 28(5), 777-795.

[10] Creal, D.D., S.J.Koopman, A. Lucas, and M. Zamojski, 2015, Generalized Autoregressive Method of Moments, Tinbergen Institute Discussion Paper, TI 2015-138/III.

[11] Diebold, F.X. and R.S. Mariano, 1995. Comparing predictive accuracy, *Journal of Business & Economic Statistics,*13(3), 253–263.

[12] Dimitriadis, T. and S. Bayer, 2017, A Joint Quantile and Expected Shortfall Regression Framework, working paper, available at `arXiv:1704.02213v1`.

[13] Dobrev, D., T.D. Nesmith and D.H. Oh, 2017, Accurate Evaluation of Expected Shortfall for Linear Portfolios with Elliptically Distributed Risk Factors, *Journal of Risk and Financial Management*, 10(5), 1-14.

[14] Engle, R.F. and S. Manganelli, 2004a, CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles, *Journal of Business & Economic Statistics*, 22, 367-381.

[15] Engle, R.F. and S. Manganelli, 2004b, A Comparison of Value-at-Risk Models in Finance, in Giorgio Szego (ed.) *Risk Measures for the 21st Century*, Wiley.

[16] Engle, R.F. and J.R. Russell, 1998, Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data, *Econometrica*, 66, 1127-1162.

[17] Fissler, T., 2017, *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Weiner Space*, PhD thesis, University of Bern.

[18] Fissler, T., and J. F. Ziegel, 2016, Higher order elicitability and Osband's principle, *Annals of Statistics*, 44(4), 1680-1707.

[19] Francq, C. and J.-M. Zakoïan, 2010, *GARCH Models*, John Wiley & Sons, United Kingdom.

[20] Francq, C. and J.-M. Zakoïan, 2015, Risk-parameter estimation in volatility models, *Journal of Econometrics*, 184, 158-173.

[21] Gerlach, R. and C.W.S. Chen, 2015, Bayesian Expected Shortfall Forecasting Incorporating the Intraday Range, *Journal of Financial Econometrics*, 14(1), 128-158.

[22] Gneiting, T., 2011, Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, 106(494), 746-762.

[23] Gschöpf, P., W.K. Härdle, and A. Mihoci, Tail Event Risk Expectile based Shortfall, SFB 649 Discussion Paper 2015-047.

[24] Hall, P., and C. C. Heyde, 1980, *Martingale Limit Theory and Its Application*, Academic Press, New York.

[25] Hansen, B.E., 1994, Autoregressive Conditional Density Estimation, *International Economic Review*, 35(3), 705-730.

[26] Harvey, A.C., 2013, *Dynamic Models for Volatility and Heavy Tails*, Econometric Society Monograph 52, Cambridge University Press, Cambridge.

[27] Huber, P.J., 1967, The behavior of maximum likelihood estimates under nonstandard conditions, in (ed.s) L.M. Le Cam and J. Neyman *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley.

[28] Jondeau, E. and M. Rockinger, 2003, Conditional volatility, skewness, and kurtosis: Existence, persistence, and comovements, *Journal of Economic Dynamics & Control*, 27, 1699-1737.

[29] Komunjer, I., 2005, Quasi-Maximum Likelihood Estimation for Conditional Quantiles, *Journal of Econometrics*, 128(1), 137-164.

[30] Komunjer, I., 2013, Quantile Prediction, in (ed.s) G. Elliott, and A. Timmermann, *Handbook of Economic Forecasting*, Vol. 2. Elsevier, Oxford.

[31] Koopman, S.J., A. Lucas and M. Scharth, Predicting Time-Varying Parameters with Parameter Driven and Observation-Driven Models, *Review of Economics and Statistics*, 98(1), 97-110.

[32] Newey, W.K. and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, in R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics*, Vol. IV, Elsevier.

[33] Newey, W.K. and J.L. Powell, 1987, Asymmetric least squares estimation and testing, *Econometrica*, 55(4), 819-847.

[34] Nolde, N. and J. F. Ziegel, 2017, Elicitability and backtesting: Perspectives for banking regulation, *Annals of Applied Statistics*, forthcoming.

[35] Patton, A.J., 2011, Volatility Forecast Comparison using Imperfect Volatility Proxies, *Journal of Econometrics*, 160(1), 246-256.

[36] Patton, A.J., 2016, Comparing Possibly Misspecified Forecasts, working paper, Duke University.

[37] Patton, A.J. and K. Sheppard, 2009, Evaluating Volatility and Correlation Forecasts, in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.) *Handbook of Financial Time Series*, Springer Verlag.

[38] Pötscher, B.M. and I.R. Prucha, 1989, A uniform law of large numbers for dependent and heterogeneous data processes, *Econometrica*, 57, 675–683.

[39] Taylor, J.W., 2008, Estimating Value-at-Risk and Expected Shortfall using Expectiles, *Journal of Financial Econometrics*, 231-252.

[40] Taylor, J.W., 2017, Forecasting Value at Risk and Expected Shortfall using a Semiparametric Approach Based on the Asymmetric Laplace Distribution, *Journal of Business & Economic Statistics*, forthcoming.

[41] Weiss, A.A., 1991, Estimating Nonlinear Dynamic Models Using Least Absolute Error Estimation, *Econometric Theory*, 7(1), 46-68.

[42] White, H. 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No. 22, Cambridge University Press.

[43] Zhu, D. and J.W. Galbraith, 2011, Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions, *Journal of Empirical Finance*, 18, 765-778.

[44] Zwingmann T. and H. Holzmann, 2016, Asymptotics for Expected Shortfall, working paper, available at `arXiv:1611.07222`.

Table 1: Simulation results for Normal innovations

	$T = 2500$				$T = 5000$			
	β	γ	b_α	c_α	β	γ	b_α	c_α
				$\alpha = 0.01$				
True	0.900	0.050	-2.665	0.873	0.900	0.050	-2.665	0.873
Median	0.901	0.049	-2.615	0.882	0.899	0.049	-2.671	0.877
Avg bias	-0.017	0.015	-0.108	0.008	-0.011	0.006	-0.089	0.004
St dev	0.077	0.076	1.095	0.022	0.049	0.033	0.805	0.015
Coverage	0.868	0.827	0.875	0.919	0.884	0.876	0.888	0.937
				$\alpha = 0.025$				
True	0.900	0.050	-2.338	0.838	0.900	0.050	-2.338	0.838
Median	0.899	0.047	-2.329	0.842	0.897	0.048	-2.392	0.841
Avg bias	-0.017	0.007	-0.137	0.004	-0.011	0.002	-0.111	0.002
St dev	0.066	0.044	0.852	0.017	0.050	0.024	0.656	0.012
Coverage	0.898	0.870	0.911	0.931	0.912	0.888	0.925	0.923
				$\alpha = 0.05$				
True	0.900	0.050	-2.063	0.797	0.900	0.050	-2.063	0.797
Median	0.901	0.048	-2.051	0.800	0.899	0.049	-2.094	0.799
Avg bias	-0.013	0.005	-0.097	0.002	-0.008	0.002	-0.081	0.001
St dev	0.062	0.046	0.707	0.015	0.041	0.021	0.511	0.010
Coverage	0.913	0.874	0.916	0.947	0.923	0.907	0.927	0.948
				$\alpha = 0.10$				
True	0.900	0.050	-1.755	0.730	0.900	0.050	-1.755	0.730
Median	0.900	0.048	-1.769	0.730	0.898	0.048	-1.778	0.730
Avg bias	-0.015	0.006	-0.103	0.000	-0.009	0.001	-0.072	0.000
St dev	0.065	0.052	0.623	0.013	0.040	0.020	0.435	0.009
Coverage	0.917	0.883	0.925	0.954	0.922	0.902	0.934	0.960
				$\alpha = 0.20$				
True	0.900	0.050	-1.400	0.601	0.900	0.050	-1.400	0.601
Median	0.898	0.048	-1.391	0.602	0.899	0.048	-1.417	0.602
Avg bias	-0.017	0.008	-0.091	0.000	-0.010	0.002	-0.064	0.000
St dev	0.078	0.072	0.547	0.014	0.044	0.022	0.374	0.010
Coverage	0.925	0.881	0.934	0.948	0.941	0.923	0.945	0.954

Notes: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with standard Normal innovations. Details are described in Section 4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 2: Simulation results for skew t innovations

	$T = 2500$				$T = 5000$			
	β	γ	b_α	c_α	β	γ	b_α	c_α
				$\alpha = 0.01$				
True	0.900	0.050	-4.506	0.730	0.900	0.050	-4.506	0.730
Median	0.893	0.049	-4.376	0.750	0.895	0.048	-4.562	0.741
Avg bias	-0.047	0.038	-0.399	0.018	-0.028	0.014	-0.340	0.009
St dev	0.150	0.134	2.687	0.048	0.094	0.065	1.983	0.034
Coverage	0.797	0.797	0.809	0.894	0.837	0.853	0.839	0.936
				$\alpha = 0.025$				
True	0.900	0.050	-3.465	0.695	0.900	0.050	-3.465	0.695
Median	0.895	0.047	-3.448	0.705	0.896	0.048	-3.520	0.701
Avg bias	-0.028	0.014	-0.254	0.008	-0.017	0.005	-0.198	0.004
St dev	0.101	0.069	1.591	0.034	0.068	0.033	1.192	0.023
Coverage	0.855	0.835	0.877	0.921	0.874	0.893	0.887	0.939
				$\alpha = 0.05$				
True	0.900	0.050	-2.767	0.651	0.900	0.050	-2.767	0.651
Median	0.896	0.048	-2.760	0.656	0.898	0.048	-2.795	0.654
Avg bias	-0.021	0.007	-0.187	0.005	-0.011	0.003	-0.114	0.003
St dev	0.081	0.049	1.085	0.025	0.053	0.025	0.782	0.017
Coverage	0.906	0.883	0.921	0.937	0.916	0.904	0.922	0.951
				$\alpha = 0.10$				
True	0.900	0.050	-2.122	0.577	0.900	0.050	-2.122	0.577
Median	0.897	0.048	-2.121	0.579	0.898	0.048	-2.140	0.578
Avg bias	-0.017	0.006	-0.125	0.003	-0.008	0.002	-0.069	0.002
St dev	0.066	0.045	0.745	0.020	0.040	0.022	0.510	0.014
Coverage	0.931	0.900	0.937	0.949	0.926	0.925	0.927	0.947
				$\alpha = 0.20$				
True	0.900	0.050	-1.514	0.431	0.900	0.050	-1.514	0.431
Median	0.899	0.050	-1.485	0.432	0.899	0.049	-1.503	0.432
Avg bias	-0.019	0.006	-0.089	0.001	-0.008	0.002	-0.049	0.001
St dev	0.089	0.047	0.618	0.018	0.042	0.022	0.380	0.012
Coverage	0.916	0.888	0.922	0.938	0.929	0.916	0.940	0.944

Notes: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with skew t innovations. Details are described in Section 4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 3: Simulation results for T=500

	Normal innovations				*Skewed t innovations*			
	β	γ	b_α	c_α	β	γ	b_α	c_α
				$\alpha = 0.01$				
True	0.900	0.050	-2.665	0.873	0.900	0.050	-4.506	0.730
Median	0.915	0.048	-2.165	0.917	0.906	0.033	-3.694	0.813
Avg bias	-0.041	0.063	0.056	0.042	-0.086	0.096	-0.250	0.071
St dev	0.161	0.189	1.550	0.049	0.233	0.264	3.552	0.095
Coverage	0.781	0.709	0.779	0.730	0.704	0.666	0.747	0.762
				$\alpha = 0.025$				
True	0.900	0.050	-2.338	0.838	0.900	0.050	-3.465	0.695
Median	0.909	0.044	-2.136	0.860	0.906	0.030	-3.170	0.736
Avg bias	-0.028	0.031	-0.048	0.020	-0.053	0.048	-0.262	0.037
St dev	0.134	0.134	1.205	0.039	0.176	0.192	2.240	0.070
Coverage	0.862	0.765	0.868	0.899	0.817	0.717	0.835	0.875
				$\alpha = 0.05$				
True	0.900	0.050	-2.063	0.797	0.900	0.050	-2.767	0.651
Median	0.905	0.040	-1.976	0.808	0.899	0.028	-2.671	0.672
Avg bias	-0.030	0.027	-0.133	0.011	-0.053	0.028	-0.366	0.021
St dev	0.134	0.142	1.098	0.033	0.174	0.165	1.817	0.053
Coverage	0.870	0.749	0.870	0.922	0.829	0.712	0.862	0.920
				$\alpha = 0.10$				
True	0.900	0.050	-1.755	0.730	0.900	0.050	-2.122	0.577
Median	0.902	0.038	-1.694	0.736	0.897	0.031	-2.195	0.588
Avg bias	-0.032	0.024	-0.156	0.006	-0.058	0.025	-0.373	0.012
St dev	0.132	0.137	0.970	0.029	0.175	0.164	1.413	0.045
Coverage	0.883	0.756	0.880	0.950	0.845	0.746	0.876	0.922
				$\alpha = 0.20$				
True	0.900	0.050	-1.400	0.601	0.900	0.050	-1.514	0.431
Median	0.899	0.037	-1.394	0.602	0.894	0.033	-1.564	0.435
Avg bias	-0.042	0.036	-0.177	0.001	-0.063	0.027	-0.290	0.004
St dev	0.147	0.174	0.854	0.031	0.187	0.167	1.070	0.038
Coverage	0.894	0.757	0.888	0.944	0.853	0.741	0.866	0.955

Notes: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with standard Normal innovations (left panel) or skew t innovations (right panel). Details are described in Section 4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 4: Sampling variation of FZ estimation relative to (Q)MLE and CAViaR

	Normal innovations				*Skew t innovations*			
	$T = 2500$		$T = 5000$		$T = 2500$		$T = 5000$	
α	β	γ	β	γ	β	γ	β	γ
Panel A: FZ/(Q)ML								
0.01	1.209	5.940	1.701	3.731	1.577	4.830	2.533	3.723
0.025	1.034	3.394	1.764	2.694	1.055	2.485	1.853	1.905
0.05	0.980	3.576	1.431	2.377	0.850	1.784	1.426	1.458
0.10	1.021	4.074	1.406	2.302	0.698	1.627	1.095	1.250
0.20	1.224	5.558	1.543	2.497	0.939	1.710	1.145	1.242
Panel B: FZ/CAViaR								
0.01	0.982	1.162	0.951	0.975	1.062	1.384	0.912	1.465
0.025	0.965	1.139	0.971	1.042	0.976	1.030	0.974	0.997
0.05	0.925	1.238	0.910	0.930	0.885	0.819	0.920	0.903
0.10	0.940	1.283	0.847	0.827	0.831	0.903	0.816	0.819
0.20	0.855	0.671	0.703	0.510	0.736	0.437	0.503	0.515

Notes: This table presents the ratio of cross-simulation standard deviations of parameter estimates obtained by FZ loss minimization and (Q)MLE (Panel A), and CAViaR (Panel B). We consider only the parameters that are common to these three estimation methods, namely the GARCH(1,1) parameters β and γ. Ratios greater than one indicate the FZ estimator is more variable than the alternative estimation method; ratios less than one indicate the opposite.

Notes to Table 5 (next page): This table presents results on the accuracy of the fitted VaR and ES estimates for the three estimation methods: QML, CAViaR and FZ estimation. In the first column of each panel we present the mean absolute error (MAE) from QML, computed across all dates in a given sample and all 1000 simulation replications. The next two columns present the *relative* MAE of CAViaR and FZ to QML. Values greater than one indicate QML is more accurate (has lower MAE); values less than one indicate the opposite.

Table 5: Mean absolute errors for VaR and ES estimates

	T = 2500						*T = 5000*					
	VaR			**ES**			**VaR**			**ES**		
	MAE	*MAE ratio*		*MAE*	*MAE ratio*		*MAE*	*MAE ratio*		*MAE*	*MAE ratio*	
α	QML	CAViaR	FZ	QML	CAViaR	FZ	QML	CAViaR	FZ	QML	CAViaR	FZ
Panel A: *Normal innovations*												
0.01	0.069	1.368	1.369	0.084	1.487	1.345	0.049	1.404	1.387	0.060	1.443	1.344
0.025	0.055	1.305	1.288	0.064	1.341	1.290	0.038	1.306	1.291	0.044	1.348	1.313
0.05	0.043	1.302	1.271	0.051	1.332	1.289	0.031	1.314	1.264	0.036	1.350	1.290
0.10	0.034	1.322	1.253	0.042	1.394	1.302	0.024	1.365	1.265	0.029	1.449	1.320
0.20	0.026	1.443	1.257	0.033	1.652	1.377	0.018	1.458	1.241	0.023	1.706	1.377
Panel B: *Skew t innovations*												
0.01	0.196	1.327	1.381	0.342	1.249	1.252	0.138	1.369	1.375	0.245	1.256	1.248
0.025	0.120	1.228	1.244	0.205	1.166	1.166	0.087	1.245	1.234	0.145	1.197	1.185
0.05	0.084	1.193	1.166	0.141	1.154	1.129	0.061	1.184	1.143	0.101	1.164	1.119
0.10	0.056	1.168	1.089	0.098	1.160	1.083	0.041	1.155	1.067	0.071	1.158	1.069
0.20	0.034	1.301	1.087	0.066	1.404	1.121	0.024	1.316	1.066	0.048	1.409	1.089
Panel C: *Offsetting dynamics in VaR and ES*												
0.01	0.111	0.395	0.406	0.264	1.048	0.420	0.075	0.278	0.281	0.260	0.310	0.287
0.025	0.074	0.339	0.343	0.222	0.679	0.331	0.055	0.251	0.255	0.221	0.253	0.249
0.05	0.058	0.369	0.348	0.196	0.522	0.298	0.046	0.262	0.267	0.196	0.264	0.233
0.10	0.058	0.476	0.465	0.166	0.436	0.348	0.048	0.382	0.388	0.166	0.341	0.292
0.20	0.054	0.977	1.030	0.106	0.778	0.508	0.049	0.902	1.019	0.106	0.744	0.459
Panel D: *Amplifying dynamics in VaR and ES*												
0.01	0.141	0.533	0.457	0.340	0.700	0.415	0.150	0.543	0.457	0.363	0.397	0.412
0.025	0.144	0.650	0.551	0.288	0.705	0.501	0.148	0.658	0.616	0.268	0.476	0.552
0.05	0.127	0.698	0.658	0.191	0.737	0.666	0.124	0.777	0.724	0.176	0.652	0.703
0.10	0.084	0.581	0.637	0.112	0.781	0.745	0.082	0.689	0.726	0.105	0.734	0.780
0.20	0.047	0.518	0.559	0.066	0.844	0.709	0.046	0.508	0.553	0.063	0.827	0.731

Notes: See previous page.

Table 6: Summary statistics

	S&P 500	DJIA	NIKKEI	FTSE
Mean (Annualized)	6.776	7.238	-2.682	3.987
Std dev (Annualized)	17.879	17.042	24.667	17.730
Skewness	-0.244	-0.163	-0.114	-0.126
Kurtosis	11.673	11.116	8.580	8.912
VaR-0.01	-3.118	-3.034	-4.110	-3.098
VaR-0.025	-2.324	-2.188	-3.151	-2.346
VaR-0.05	-1.731	-1.640	-2.451	-1.709
VaR-0.10	-1.183	-1.126	-1.780	-1.193
ES-0.01	-4.528	-4.280	-5.783	-4.230
ES-0.025	-3.405	-3.215	-4.449	-3.295
ES-0.05	-2.697	-2.553	-3.603	-2.643
ES-0.10	-2.065	-1.955	-2.850	-2.031

Notes: This table presents summary statistics on the four daily equity return series studied in Section 5, over the full sample period from January 1990 to December 2016. The first two rows report the annualized mean and standard deviation of these returns in percent. The second panel presents sample Value-at-Risk for four choices of α, and the third panel presents corresponding sample Expected Shortfall estimates.

Table 7: ARMA, GARCH, and Skew t results

	SP500	DJIA	NIKKEI	FTSE
ϕ_0	0.056	0.056	-0.029	0.042
θ_1	–	–	–	0.075
R^2	0.000	0.000	0.000	0.006
ω	0.005	0.004	0.072	0.009
β	0.942	0.922	0.865	0.936
γ	0.052	0.077	0.105	0.053
ν	6.358	6.766	6.677	13.663
λ	-0.035	-0.059	-0.016	-0.024

Notes: This table presents parameter estimates for the four daily equity return series studied in Section 5, over the in-sample period from January 1990 to December 1999. The first panel presents the optimal ARMA model according to the BIC, along with the R^2 of that model. The second panel presents the estimated GARCH(1,1) parameters, and the third panel presents the estimated parameters of the skewed t distribution applied to the estimated standardized residuals.

Table 8: Estimated paramters of GAS models for VaR and ES

	GAS-2F			*GAS-1F*	*GARCH-FZ*	*Hybrid*
	VaR	**ES**				
w	-0.009	-0.010	β	0.995	0.944	0.974
(s.e.)	(0.003)	(0.004)	(s.e.)	(0.002)	(0.058)	(0.006)
b	0.993	0.994	γ	0.007	0.031	0.003
(s.e.)	(0.002)	(0.003)	(s.e.)	(0.0001)	(0.010)	(0.003)
a_v	-0.358	-0.351	δ	–	–	0.017
(s.e.)	(0.109)	(0.129)	(s.e.)			(0.002)
a_e	-0.003	-0.003	a	-1.164	-1.955	-2.320
(s.e.)	(0.002)	(0.003)	(s.e.)	(0.420)	(0.256)	(4.671)
			b	-1.757	-2.829	-3.434
			(s.e.)	(0.634)	(0.522)	(6.874)
Avg loss	0.592			0.603	0.637	0.590
Time (secs)	44.773			0.594	0.767	1.343

Notes: This table presents parameter estimates and standard errors for four GAS models of VaR and ES for the S&P 500 index over the in-sample period from January 1990 to December 1999. The left panel presents the results for the two-factor GAS model in Section 2.2. The right panel presents the results for the three one-factor models: a one-factor GAS model (from Section 2.3), and a GARCH model estimated by FZ loss minimization, and "hybrid" one-factor GAS model that includes a additional GARCH-type forcing variable (both from Section 2.5). The penultimate row of this table presents the average (in-sample) losses from each of these four models, and the bottom row presents the estimation time for each model (using Matlab R2018b on a 3.4GHz machine).

Table 9: Out-of-sample average losses and goodness-of-fit tests (alpha=0.05)

	Average loss				*GoF p-values: VaR*				*GoF p-values: ES*			
	S&P	**DJIA**	**NIK**	**FTSE**	**S&P**	**DJIA**	**NIK**	**FTSE**	**S&P**	**DJIA**	**NIK**	**FTSE**
RW-125	0.914	0.864	1.290	0.959	0.039	0.021	0.002	0.000	0.046	0.028	0.011	0.000
RW-250	0.959	0.909	1.294	1.002	0.003	0.002	0.026	0.000	*0.062*	0.020	0.034	0.002
RW-500	1.023	0.975	1.318	1.056	0.002	0.003	0.001	0.000	0.024	0.021	0.002	0.000
GCH-N	0.876	0.811	1.170	0.871	0.043	0.010	**0.536**	0.001	0.001	0.000	**0.195**	0.000
GCH-Skt	0.865	0.799	1.168	*0.864*	0.006	0.006	**0.109**	0.001	0.005	0.004	**0.273**	0.000
GCH-EDF	0.862	*0.796*	*1.166*	0.865	0.005	0.006	**0.580**	0.001	0.019	0.018	**0.519**	0.000
FZ-2F	*0.859*	0.799	1.206	0.874	0.004	0.001	**0.279**	0.001	**0.170**	**0.319**	**0.313**	0.004
FZ-1F	**0.850**	**0.791**	1.190	0.871	0.007	0.000	**0.218**	0.000	*0.073*	0.002	**0.550**	0.001
GCH-FZ	0.862	0.797	1.166	0.870	0.009	0.008	**0.519**	0.000	0.027	0.035	**0.459**	0.000
Hybrid	0.870	0.796	**1.165**	**0.859**	0.000	0.007	**0.464**	0.000	0.002	0.038	**0.453**	0.000

Notes: The left panel of this table presents the average losses, using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The lowest average loss in each column is highlighted in bold, the second-lowest is highlighted in italics. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2. The middle and right panels of this table present p-values from goodness-of-fit tests of the VaR and ES forecasts respectively. Values that are greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and values between 0.05 and 0.10 are in italics.

Table 10: Diebold-Mariano t-statistics on average out-of-sample loss differences alpha=0.05, S&P 500 returns

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
RW125		-2.257	-3.527	1.952	2.478	2.625	2.790	3.600	2.736	2.684
RW250	2.257		-3.215	2.752	3.129	3.246	3.364	4.039	3.399	3.538
RW500	3.527	3.215		3.706	3.997	4.087	4.334	4.818	4.223	4.454
G-N	-1.952	-2.752	-3.706		3.526	2.965	1.418	2.483	2.847	0.634
G-Skt	-2.478	-3.129	-3.997	-3.526		1.954	0.626	1.791	1.179	-0.436
G-EDF	-2.625	-3.246	-4.087	-2.965	-1.954		0.335	1.529	-0.023	-0.756
FZ-2F	-2.790	-3.364	-4.334	-1.418	-0.626	-0.335		1.000	-0.329	-0.904
FZ-1F	-3.600	-4.039	-4.818	-2.483	-1.791	-1.529	-1.000		-1.624	-2.049
G-FZ	-2.736	-3.399	-4.223	-2.847	-1.179	0.023	0.329	1.624		-0.895
Hybrid	-2.684	-3.538	-4.454	-0.634	0.436	0.756	0.904	2.049	0.895	

Notes: This table presents t-statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2.

Table 11: Out-of-sample performance rankings for various alpha

	$\alpha = 0.01$					$\alpha = 0.025$				
	S&P	**DJIA**	**NIK**	**FTSE**	**Avg**	**S&P**	**DJIA**	**NIK**	**FTSE**	**Avg**
RW-125	8	8	10	8	8.50	8	8	9	7	8.00
RW-250	9	9	8	9	8.75	9	9	8	8	8.50
RW-500	10	10	9	10	9.75	10	10	10	10	10.00
G-N	7	7	5	4	5.75	7	6	4	3	5.00
G-Skt	6	3	1	2	3.00	5	3	1	1	2.50
G-EDF	5	2	2	1	2.50	2	2	3	2	2.25
FZ-2F	4	4	6	7	5.25	4	5	7	9	6.25
FZ-1F	3	6	7	6	5.50	3	4	6	5	4.50
G-FZ	2	1	3	3	2.25	1	1	2	4	2.00
Hybrid	1	5	4	5	3.75	6	7	5	6	6.00

	$\alpha = 0.05$					$\alpha = 0.10$				
	S&P	**DJIA**	**NIK**	**FTSE**	**Avg**	**S&P**	**DJIA**	**NIK**	**FTSE**	**Avg**
RW-125	8	8	8	8	8.00	8	8	8	8	8.00
RW-250	9	9	9	9	9.00	9	9	9	9	9.00
RW-500	10	10	10	10	10.00	10	10	10	10	10.00
G-N	7	7	5	6	6.25	3	4	7	4	4.50
G-Skt	5	6	4	2	4.25	7	6	6	3	5.50
G-EDF	3	3	2	3	2.75	4	2	3	5	3.50
FZ-2F	2	2	7	4	3.75	2	3	5	7	4.25
FZ-1F	1	1	6	7	3.75	1	7	2	2	3.00
G-FZ	4	5	3	5	4.25	6	5	4	6	5.25
Hybrid	6	4	1	1	3.00	5	1	1	1	2.00

Notes: This table presents the rankings (with the best performing model ranked 1 and the worst ranked 10) based on average losses using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The first three rows in each panel correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2. The last column in each panel represents the average rank across the four equity return series.

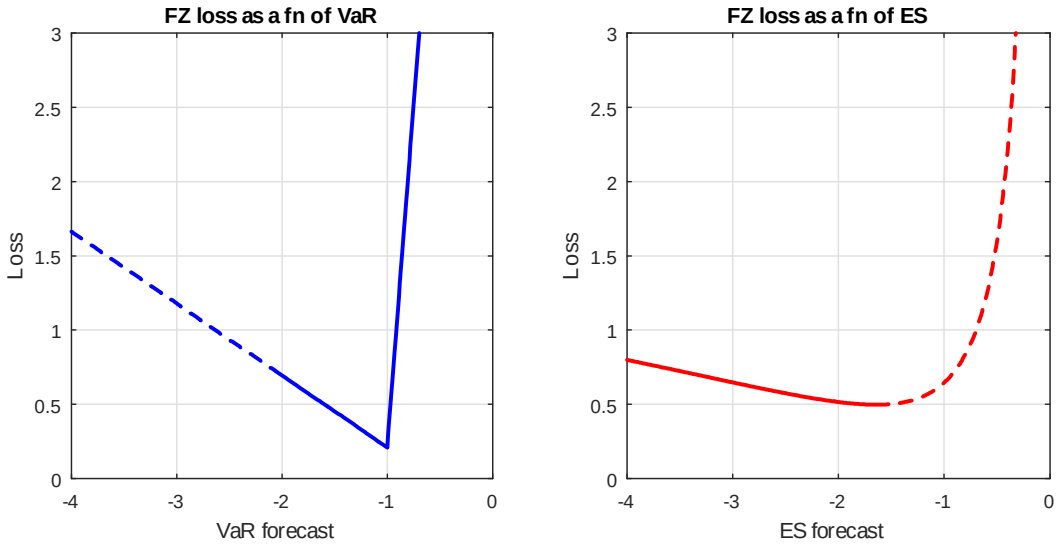

Figure 1: *This figure plots the FZ0 loss function when* $Y = -1$ *and* $\alpha = 0.05$. *In the left panel we fix* $e = -2.06$ *and vary* v, *in the right panel we fix* $v = -1.64$ *and vary* e. *Values where* $v < e$ *are indicated with a dashed line.*

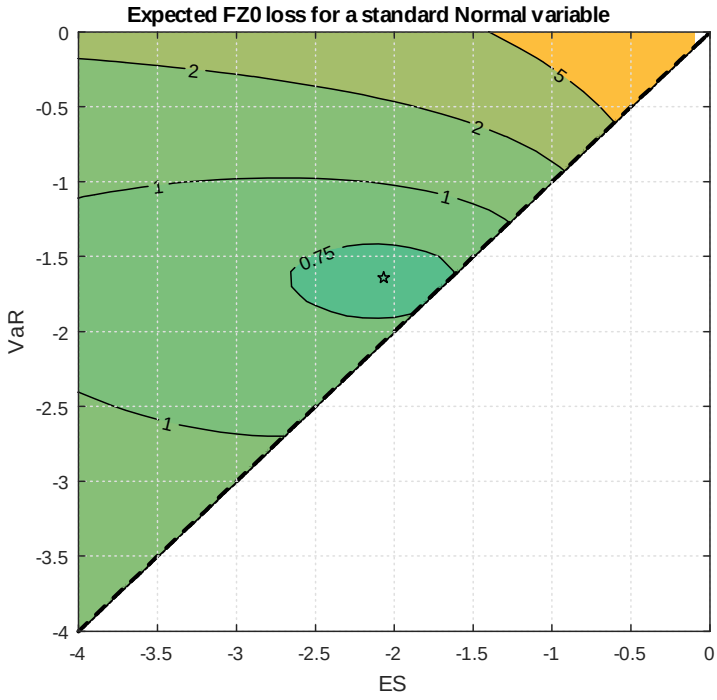

Figure 2: *Contours of expected FZ0 loss when the target variable is standard Normal. Only values where ES<VaR<0 are considered. The optimal value is marked with a star.*

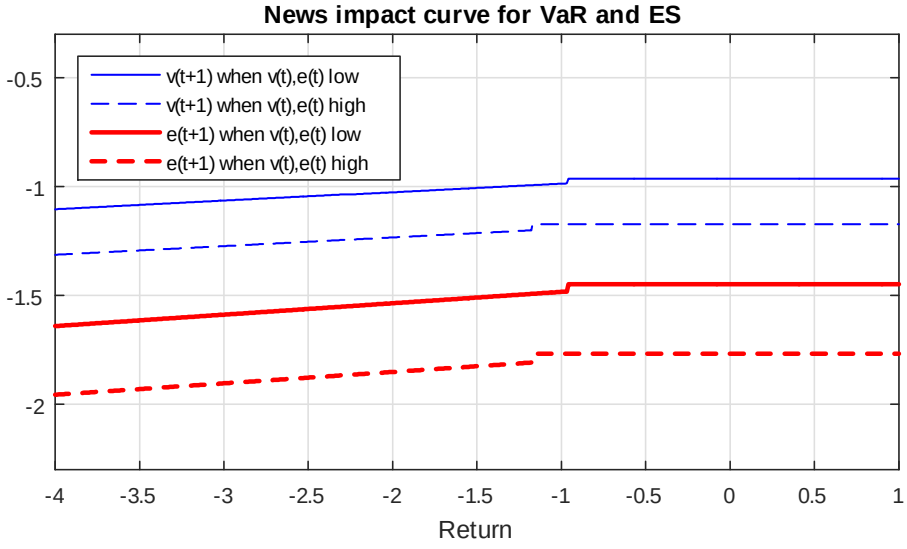

Figure 3: *This figure shows the values of VaR and ES as a function of the lagged return, when the lagged values of VaR and ES are either low (10% below average) or high (10% above average). The function is based on the estimated parameters for daily S&P 500 returns.*

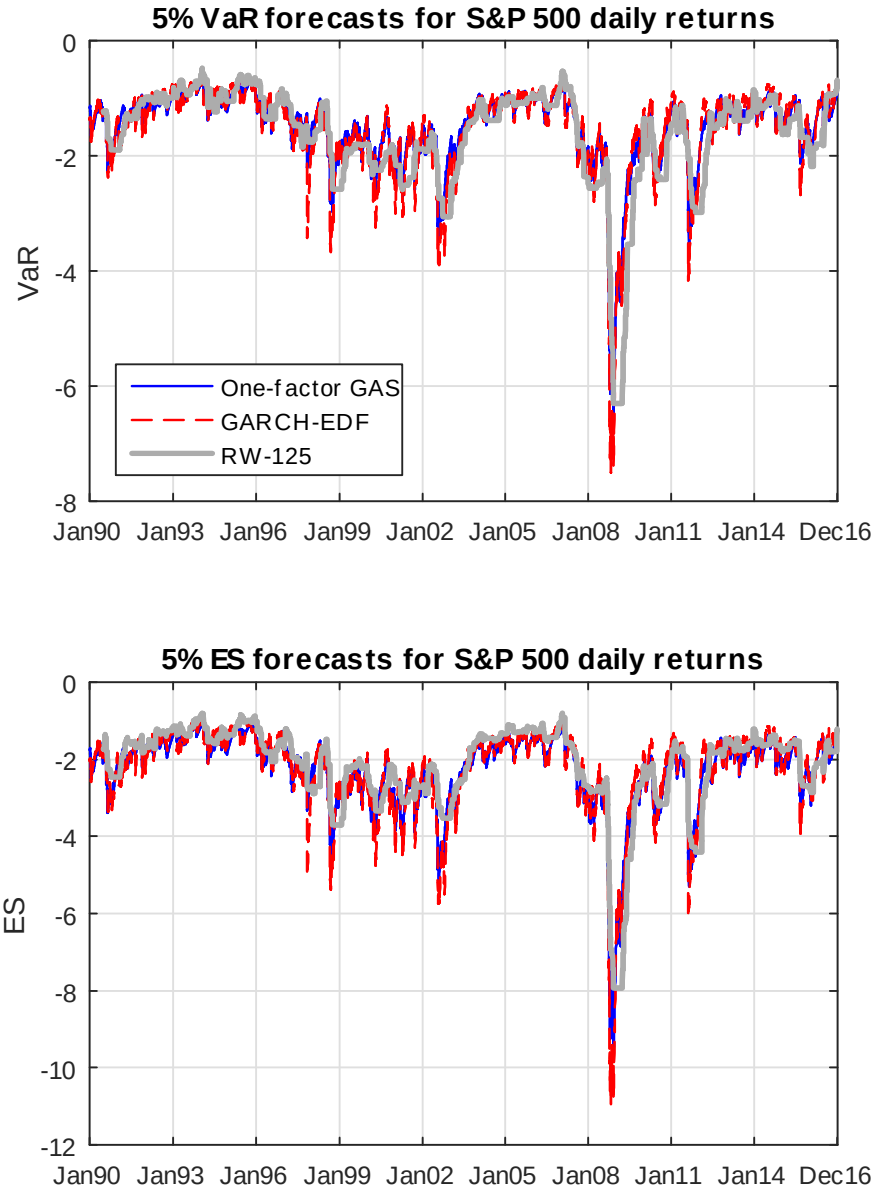

Figure 4: *This figure plots the estimated 5% Value-at-Risk (VaR) and Expected Shortfall (ES) for daily returns on the S&P 500 index, over the period January 1990 to December 2016. The estimates are based on a one-factor GAS model, a GARCH model, and a rolling window using 125 observations.*

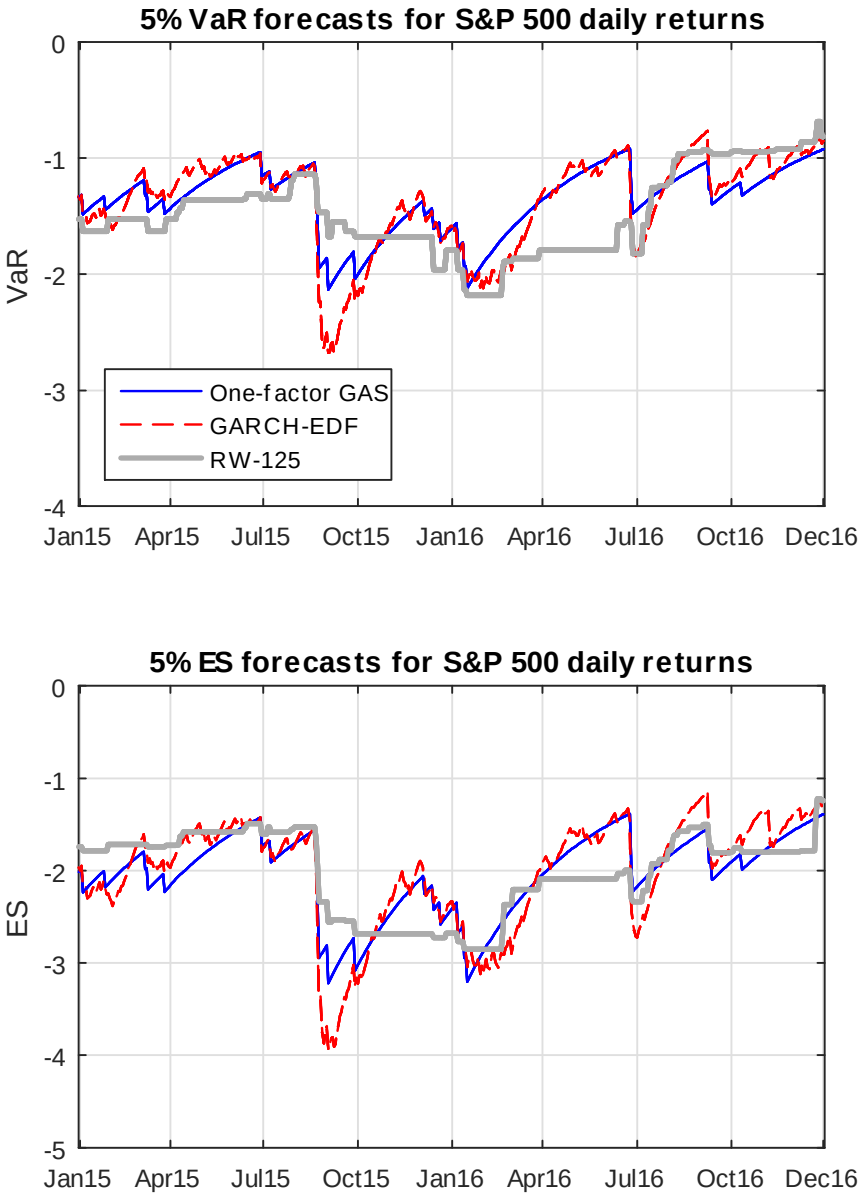

Figure 5: *This figure plots the estimated 5% Value-at-Risk (VaR) and Expected Shortfall (ES) for daily returns on the S&P 500 index, over the period January 2015 to December 2016. The estimates are based on a one-factor GAS model, a GARCH model, and a rolling window using 125 observations.*

Supplemental Appendix to:

Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)

Andrew J. Patton, Duke University
Johanna F. Ziegel, University of Bern
Rui Chen, Duke University

12 September 2018

This appendix contains three parts. Part 1 presents lemmas that provide further details on the proof of Theorem 2 presented in the main paper. Part 2 presents a detailed verification that the high level assumptions made in the theorems of the paper hold for the widely-used GARCH(1,1) process. Part 3 contains additional tables of analysis.

Appendix SA.1: Detailed proofs

Throughout this appendix, we suppress the subscript on $\hat{\boldsymbol{\theta}}_T$ for simplicity of presentation, and we denote the conditional distribution and density functions as F_t and f_t rather than $F_t\left(\cdot|\mathcal{F}_{t-1}\right)$ and $f_t\left(\cdot|\mathcal{F}_{t-1}\right)$.

In Lemmas 1 and 3 below, we will refer to the expected score, defined as:

$$\begin{aligned}\lambda(\boldsymbol{\theta}) &= \mathbb{E}\left[g_t(\boldsymbol{\theta})\right] \\ &= \mathbb{E}\left[\frac{1}{-e_t(\boldsymbol{\theta})}\left(\frac{F_t\left(v_t(\boldsymbol{\theta})\right)}{\alpha}-1\right)\nabla v_t(\boldsymbol{\theta})' + \right. \\ &\left. \frac{1}{e_t(\boldsymbol{\theta})^2}\left(\frac{F_t\left(v_t(\boldsymbol{\theta})\right)}{\alpha}v_t(\boldsymbol{\theta}) - \frac{1}{\alpha}\mathbb{E}_{t-1}[Y_t|1\left\{Y_t \leq v_t(\boldsymbol{\theta})\right\}] - v_t(\boldsymbol{\theta}) + e_t(\boldsymbol{\theta})\right)\nabla e_t(\boldsymbol{\theta})'\right]\end{aligned} \tag{1}$$

Lemma 1 *Let*

$$\Lambda(\boldsymbol{\theta}^*) = \left.\frac{\partial \mathbb{E}\left[g_t(\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{2}$$

Then under Assumptions 1-2,

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \left(\Lambda^{-1}(\boldsymbol{\theta}^0) + o_p(1)\right)\left(-\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^0) + o_p(1)\right) \tag{3}$$

Proof of Lemma 1. Consider a mean-value expansion of $\lambda(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^0$:

$$\lambda(\hat{\boldsymbol{\theta}}) = \lambda(\boldsymbol{\theta}^0) + \left.\frac{\partial \mathbb{E}\left[g_t(\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \tag{4}$$

$$= \Lambda(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \tag{5}$$

where $\boldsymbol{\theta}^*$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$, and noting that $\lambda(\boldsymbol{\theta}^0) = 0$ and the definition of $\Lambda(\boldsymbol{\theta}^*)$ given in the statement of the lemma. Proving the claim involves two results: (I) $\Lambda^{-1}(\boldsymbol{\theta}^*) = \Lambda^{-1}(\boldsymbol{\theta}^0) + o_p(1)$, and (II) $\sqrt{T}\lambda(\hat{\boldsymbol{\theta}}) = -\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^0) + o_p(1)$. Part (I) is easy to verify: Since $v_t(\boldsymbol{\theta})$ and $e_t(\boldsymbol{\theta})$ are twice continuously differentiable, and $e_t(\boldsymbol{\theta}^0) < 0$, $\Lambda(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\Lambda(\boldsymbol{\theta})$ is non-singular in a neighborhood of $\boldsymbol{\theta}^0$. Then by the continuous mapping theorem, $\boldsymbol{\theta}^* \xrightarrow{p} \boldsymbol{\theta}^0 \Rightarrow \Lambda(\boldsymbol{\theta}^*)^{-1} \xrightarrow{p} \Lambda^{-1}(\boldsymbol{\theta}^0)$. Establishing (II) builds on Theorem 3 of Huber (1967) and Lemma A.1 of Weiss (1991), which extends Huber's conclusion to the case of non-*iid* dependent random variables. We are going to verify the conditions of Weiss's Lemma A.1. Since the other conditions are easily checked, we only need to show that $T^{-1/2}\sum_{t=1}^{T} g_t(\hat{\boldsymbol{\theta}}) = o_p(1)$, which we show in Lemma 2, and that his assumptions N3 and N4 hold, which we show in Lemmas 3-6. ■

Lemma 2 *Under Assumptions 1-2,* $T^{-1/2}\sum_{t=1}^{T} g_t(\hat{\boldsymbol{\theta}}) = o_p(1)$.

Proof of Lemma 2. Let $\{e_j\}_{j=1}^{p}$ be the standard basis of $\mathbb{R}^p$ and define

$$L_T^j(a) = T^{-1/2}\sum_{t=1}^{T} L_{FZ0}\left(Y_t, v_t(\hat{\boldsymbol{\theta}} + ae_j), e_t(\hat{\boldsymbol{\theta}} + ae_j); \alpha\right) \tag{6}$$

where a is a scalar. Let $G_T^j(a)$ (a scalar) be the right partial derivative of $L_T^j(a)$, that is

$$G_T^j(a) = T^{-1/2}\sum_{t=1}^{T}\left(\frac{\nabla_j v_t(\hat{\boldsymbol{\theta}} + ae_j)}{-e_t(\hat{\boldsymbol{\theta}} + ae_j)}(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \le v_t(\hat{\boldsymbol{\theta}} + ae_j)\right\} - 1) + \right. \tag{7}$$

$$\left.\frac{\nabla_j e_t(\hat{\boldsymbol{\theta}} + ae_j)}{e_t(\hat{\boldsymbol{\theta}} + ae_j)^2}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \le v_t(\hat{\boldsymbol{\theta}} + ae_j)\right\}(v_t(\hat{\boldsymbol{\theta}} + ae_j) - Y_t) - v_t(\hat{\boldsymbol{\theta}} + ae_j) + e_t(\hat{\boldsymbol{\theta}} + ae_j)\right)\right)$$

$G_T^j(0) = \lim_{\xi_1 \to 0+} G_T^j(\xi_1)$ is the right partial derivative of $L_T(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$ in the direction θ_j, while $\lim_{\xi_2 \to 0+} G_T^j(-\xi_2)$ is the left partial derivative of $L_T(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$ in the direction θ_j. Because $L_T(\boldsymbol{\theta})$ achieves its minimum at $\hat{\boldsymbol{\theta}}$, and its left and right partial derivatives exist, its left derivative must

be non-positive and its right derivative must be non-negative. Thus,

$$\begin{aligned}|G_T^j(0)| &\leq \lim_{\xi_1 \to 0+} G_T^j(\xi_1) - \lim_{\xi_2 \to 0+} G_T^j(-\xi_2) \\ &= T^{-1/2} \sum_{t=1}^{T} \left(\frac{\nabla_j v_t(\hat{\boldsymbol{\theta}})}{-e_t(\hat{\boldsymbol{\theta}})} \frac{1}{\alpha} \mathbf{1}\left\{ Y_t = v_t(\hat{\theta}) \right\} + \frac{\nabla_j e_t(\hat{\boldsymbol{\theta}})}{e_t(\hat{\boldsymbol{\theta}})^2} \frac{1}{\alpha} \left(v_t(\hat{\boldsymbol{\theta}}) - Y_t \right) \mathbf{1}\left\{ Y_t = v_t(\hat{\boldsymbol{\theta}}) \right\} \right) \quad (8) \\ &= T^{-1/2} \sum_{t=1}^{T} \frac{|\nabla_j v_t(\hat{\boldsymbol{\theta}})|}{-e_t(\hat{\theta})} \frac{1}{\alpha} \mathbf{1}\left\{ Y_t = v_t(\hat{\boldsymbol{\theta}}) \right\}\end{aligned}$$

The second term in the penultimate line vanishes as $\mathbf{1}\{Y_t = v_t(\hat{\boldsymbol{\theta}})\}(v_t(\hat{\boldsymbol{\theta}}) - Y_t)$ is always zero.

By Assumption 2(C), for all t, $|\nabla_j v_t(\hat{\boldsymbol{\theta}})| \leq \|\nabla v_t(\hat{\boldsymbol{\theta}})\| \leq V_1(\mathcal{F}_{t-1})$ and $\left| 1/e_t(\hat{\boldsymbol{\theta}}) \right| \leq H$, thus:

$$|G_T^j(0)| \leq \frac{H}{\alpha} \left[T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) \right] \left[\sum_{t=1}^{T} \mathbf{1}\left\{ Y_t = v_t(\hat{\boldsymbol{\theta}}) \right\} \right] \quad (9)$$

H is finite by Assumption 2(C). Next note that for all $\epsilon > 0$,

$$\Pr\left[T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) > \epsilon \right] \leq \sum_{t=1}^{T} \Pr\left[V_1(\mathcal{F}_{t-1}) > \epsilon T^{1/2} \right] \leq \sum_{t=1}^{T} \frac{\mathbb{E}[V_1(\mathcal{F}_{t-1})^3]}{\epsilon^3 T^{3/2}} \to 0 \quad (10)$$

with the latter inequality following from Markov's inequality. Since $\mathbb{E}[V_1(\mathcal{F}_{t-1})^3]$ is finite by assumption 2(D), we then have that $T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) = o_p(1)$. Finally, by Assumption 2(G) we have $\sum_{t=1}^{T} \mathbf{1}\left\{ Y_t = v_t(\hat{\boldsymbol{\theta}}) \right\} = \mathcal{O}_{a.s.}(1)$. We therefore have $G_T^j(0) \xrightarrow{p} 0$. Since this holds for every j, we have $T^{-1/2} \sum_{t=1}^{T} g_t(\hat{\boldsymbol{\theta}}) = o_p(1)$.

■

The following three lemmas show each of the three parts of Assumption N3 of Weiss (1991) holds. In the proofs below we make repeated use of mean-value expansions, and we use $\boldsymbol{\theta}^*$ to denote a point on the line connecting $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$, and $\boldsymbol{\theta}^{**}$ to denote a point on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^0$. The particular point on the line can vary from expansion to expansion.

Lemma 3 *Under assumptions 1-2, Assumption N3(i) of Weiss (1991) holds:*

$$\|\lambda_T(\boldsymbol{\theta})\| \geq a\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\|, \text{ for } \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \leq d_0.$$

for T sufficiently large, where a and d_0 are strictly positive numbers.

Proof of Lemma 3. A mean-value expansion yields:

$$\lambda_T(\hat{\boldsymbol{\theta}}) = \lambda_T(\boldsymbol{\theta}^0) + \Lambda(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \Lambda_T(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \tag{11}$$

since $\lambda_T(\boldsymbol{\theta}^0) = 0$, where $\Lambda(\boldsymbol{\theta}) = \partial \mathbb{E}[g_t(\boldsymbol{\theta})]/\partial \boldsymbol{\theta}$. Using the fact that

$$\frac{\partial \mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\}|\mathcal{F}_{t-1}]}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \int_{-\infty}^{v_t(\boldsymbol{\theta})} y f_t(y) dy \right\} = v_t(\boldsymbol{\theta}) f_t(v_t(\boldsymbol{\theta})) \nabla v_t(\boldsymbol{\theta}) \tag{12}$$

we can write:

$$\begin{aligned}
\Lambda(\boldsymbol{\theta}) = \mathbb{E}\Bigg[& \left(\frac{\nabla^2 v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} + \frac{\nabla v_t(\boldsymbol{\theta})' \nabla e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} + \frac{\nabla e_t(\boldsymbol{\theta})' \nabla v_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right) \left(\frac{F_t(v_t(\boldsymbol{\theta}))}{\alpha} - 1 \right) \\
& + \left(\nabla^2 e_t(\boldsymbol{\theta}) \frac{1}{e_t(\boldsymbol{\theta})^2} + \frac{-2}{e_t(\boldsymbol{\theta})^3} \nabla e_t(\boldsymbol{\theta})' \nabla e_t(\boldsymbol{\theta}) \right) \\
& \cdot \left(\left(\frac{F_t(v_t(\boldsymbol{\theta}))}{\alpha} - 1 \right) v_t(\boldsymbol{\theta}) - \frac{1}{\alpha} \mathbb{E}\left[Y_t \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} | \mathcal{F}_{t-1} \right] + e_t(\boldsymbol{\theta}) \right) \\
& + \frac{f_t(v_t(\boldsymbol{\theta}))}{-\alpha e_t(\boldsymbol{\theta})} \nabla' v_t(\boldsymbol{\theta}) \nabla v_t(\boldsymbol{\theta}) \\
& + \frac{1}{e_t(\boldsymbol{\theta})^2} \nabla' e_t(\boldsymbol{\theta}) \nabla e_t(\boldsymbol{\theta})] \} \Bigg| \mathcal{F}_{t-1} \Bigg]
\end{aligned} \tag{13}$$

Evaluated at $\boldsymbol{\theta}^0$, the first two terms of Λ drop out because $F_t\left(v_t(\boldsymbol{\theta}^0)\right) = \alpha$ and $\frac{1}{\alpha}\mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta}^0)\}|\mathcal{F}_{t-1}] = e_t\left(\boldsymbol{\theta}^0\right)$. Define D as

$$D \equiv \Lambda(\boldsymbol{\theta}^0) = T^{-1} \sum_{t=1}^{T} \mathbb{E}\left[\frac{f_t(v_t(\boldsymbol{\theta}^0))}{-\alpha e_t(\boldsymbol{\theta}^0)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^0) + \frac{1}{e_t(\boldsymbol{\theta}^0)^2} \nabla e_t(\boldsymbol{\theta}^0)' \nabla e_t(\boldsymbol{\theta}^0) \right] \tag{14}$$

Below we show that $\Lambda(\boldsymbol{\theta}^*) = D + O(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|)$ by decomposing $\|\Lambda_T(\boldsymbol{\theta}^*) - D\|$ into four terms and showing that each is bounded by a $O(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|)$ term.

First term: Using a mean-value expansion around $\boldsymbol{\theta}^0$ and Assumptions 2(C)-(D) we obtain:

$$\begin{aligned}
& \left\| \mathbb{E}\left[\left(\frac{\nabla^2 v_t(\boldsymbol{\theta}^*)}{-e_t(\boldsymbol{\theta}^*)} + \frac{\nabla v_t(\boldsymbol{\theta}^*)' \nabla e_t(\boldsymbol{\theta}^*)}{e_t(\boldsymbol{\theta}^*)^2} + \frac{\nabla e_t(\boldsymbol{\theta}^*)' \nabla v_t(\boldsymbol{\theta}^*)}{e_t(\boldsymbol{\theta}^*)^2} \right) \left(\frac{F_t(v_t(\boldsymbol{\theta}^*))}{\alpha} - 1 \right) \right] \right\| \\
\leq \; & \mathbb{E}\left[\left\| \left(H V_2(\mathcal{F}_{t-1}) + 2H^2 V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1}) \right) \left(\frac{f_t(v_t(\boldsymbol{\theta}^{**}))}{\alpha} \nabla v_t(\boldsymbol{\theta}^{**})(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0) \right) \right\| \right] \\
\leq \; & \frac{K}{\alpha} \left\{ H \mathbb{E}[V_1(\mathcal{F}_{t-1})^3]^{1/3} \mathbb{E}[V_2(\mathcal{F}_{t-1})^{3/2}]^{2/3} + 2H^2 \mathbb{E}[V_1(\mathcal{F}_{t-1})^3]^{2/3} \mathbb{E}[H_1(\mathcal{F}_{t-1})^3]^{1/3} \right\} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}^0\|
\end{aligned} \tag{15}$$

Second term: Again using a mean-value expansion around $\boldsymbol{\theta}^0$ and Assumptions 2(C)-(D):

$$
\begin{aligned}
&\left\| \mathbb{E}\left[\left(\frac{1}{e_t(\boldsymbol{\theta}^*)^2} \nabla^2 e_t(\boldsymbol{\theta}^*) - \frac{2}{e_t(\boldsymbol{\theta}^*)^3} \nabla e_t(\boldsymbol{\theta}^*)' \nabla e_t(\boldsymbol{\theta}^*) \right) \right.\right. \\
&\qquad \left.\left. \cdot \left(\left(\frac{F_t(v_t(\boldsymbol{\theta}^*))}{\alpha} - 1 \right) v_t(\boldsymbol{\theta}^*) - \frac{1}{\alpha} \mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta}^*)\} | \mathcal{F}_{t-1}] + e_t(\boldsymbol{\theta}^*) \right) \right] \right\| \\
\leq & \mathbb{E}[\| \left(H_2(\mathcal{F}_{t-1}) H^2 + H_1(\mathcal{F}_{t-1}) \cdot 2H^3 \cdot H_1(\mathcal{F}_{t-1}) \right) \\
&\qquad \cdot \left(\left(F_t\left(v_t(\boldsymbol{\theta}^{**}) \right) / \alpha - 1 \right) \nabla v_t(\boldsymbol{\theta}^{**}) + \nabla e_t(\boldsymbol{\theta}^{**}) \right) (\boldsymbol{\theta}^* - \boldsymbol{\theta}^0) \|] \\
\leq & \{ (1/\alpha + 1)(H^2 \mathbb{E}[V_1(\mathcal{F}_{t-1}) H_2(\mathcal{F}_{t-1})] + 2H^3 \mathbb{E}[V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1})^2]) \\
&\qquad + (H^2 \cdot \mathbb{E}[H_1(\mathcal{F}_{t-1}) H_2(\mathcal{F}_{t-1})] + 2H^3 \mathbb{E}[H_1(\mathcal{F}_{t-1})^3]) \} \| \boldsymbol{\theta}^* - \boldsymbol{\theta}^0 \|
\end{aligned} \tag{16}
$$

Third term:

$$
\begin{aligned}
&\left\| \mathbb{E}\left[\frac{f_t(v_t(\boldsymbol{\theta}^*))}{-e_t(\theta^*)\alpha} \nabla v_t(\boldsymbol{\theta}^*)' \nabla v_t(\boldsymbol{\theta}^*) - \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)\alpha} \nabla v_t(\theta^0)' \nabla v_t(\boldsymbol{\theta}^0) \right] \right\| \\
&= \frac{1}{\alpha} \left\| T^{-1} \sum_{t=1}^{T} \mathbb{E}\{ \frac{f_t(v_t(\boldsymbol{\theta}^*))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^*)' \nabla v_t(\boldsymbol{\theta}^*) - \frac{f_t(v_t(\boldsymbol{\theta}^*))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) \right. \\
&+ \frac{f_t(v_t(\boldsymbol{\theta}^*))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) - \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) \\
&+ \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) - \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) \\
&\left. + \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) - \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^0) \} \right\| \\
&= \frac{1}{\alpha} \left\| T^{-1} \sum_{t=1}^{T} \mathbb{E}\{ \frac{f_t(v_t(\boldsymbol{\theta}^*))}{-e_t(\boldsymbol{\theta}^*)} [\nabla^2 v_t(\boldsymbol{\theta}^{**})(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0)] \nabla v_t(\theta^*) \right. \\
&+ \frac{f_t(v_t(\boldsymbol{\theta}^*)) - f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^*)} \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) \\
&+ \frac{f_t(v_t(\boldsymbol{\theta}^0))}{e_t(\boldsymbol{\theta}^{**})^2} (\boldsymbol{\theta}^* - \boldsymbol{\theta}^0) \nabla v_t(\boldsymbol{\theta}^0)' \nabla v_t(\boldsymbol{\theta}^*) \\
&\left. + \frac{f_t(v_t(\boldsymbol{\theta}^0))}{-e_t(\boldsymbol{\theta}^0)} \nabla v_t(\boldsymbol{\theta}^0)' (\boldsymbol{\theta}^* - \boldsymbol{\theta}^0)^2 v_t(\boldsymbol{\theta}^{**}) \} \right\| \\
&\leq \frac{1}{\alpha} T^{-1} \sum_{t=1}^{T} \mathbb{E}\{ V_2(\mathcal{F}_{t-1}) \left(KH \cdot V_1(\mathcal{F}_{t-1}) \right) + KH \cdot V_1(\mathcal{F}_{t-1})^3 \\
&+ KH^2 H_1(\mathcal{F}_{t-1}) V_1(\mathcal{F}_{t-1})^2 + KH V_1(\mathcal{F}_{t-1}) V_2(\mathcal{F}_{t-1}) \} \cdot \| \boldsymbol{\theta}^* - \boldsymbol{\theta}^0 \|
\end{aligned} \tag{17}
$$

Fourth term: The bound on this term follows similar steps to that of the third term:

$$
\begin{aligned}
& \left\| T^{-1}\sum_{t=1}^{T}\mathbb{E}\{\frac{1}{e_t(\boldsymbol{\theta}^*)^2}\nabla e_t(\boldsymbol{\theta}^*)'\nabla e_t(\boldsymbol{\theta}^*) - \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^0)\}\right\| \\
= & \left\| T^{-1}\sum_{t=1}^{T}\mathbb{E}\{\frac{1}{e_t(\boldsymbol{\theta}^*)^2}\nabla e_t(\boldsymbol{\theta}^*)'\nabla e_t(\boldsymbol{\theta}^*) - \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^*)\right. \\
& + \frac{1}{e_t(\boldsymbol{\theta}^*)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^*) - \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^*) \\
& \left. + \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^*) - \frac{1}{e_t(\boldsymbol{\theta}^0)^2}\nabla e_t(\boldsymbol{\theta}^0)'\nabla e_t(\boldsymbol{\theta}^0)\}\right\| | \\
\leq & T^{-1}\sum_{t=1}^{T}\{H^2\cdot\mathbb{E}[H_1(\mathcal{F}_{t-1})H_2(\mathcal{F}_{t-1})] + 2H^3\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^3\right] + H^2\mathbb{E}[H_1(\mathcal{F}_{t-1})H_2(\mathcal{F}_{t-1})]\}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^0\|
\end{aligned}
\tag{18}
$$

Therefore, $\Lambda_T(\boldsymbol{\theta}^*) = D_T + O(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|) \Rightarrow \|\Lambda_T(\boldsymbol{\theta}^*) - D_T\| \leq K\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|$,where K is some constant $< \infty$, for T sufficiently large. By Assumption 2(E), D_T has eigenvalues bounded below by a positive constant, denoted as a, for T sufficiently large. Thus,

$$
\begin{aligned}
\|\lambda_T(\hat{\boldsymbol{\theta}})\| &= \|\Lambda_T(\boldsymbol{\theta}^*)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)\| \\
&= \|D_T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) - (D_T - \Lambda_T(\boldsymbol{\theta}^*))(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\| \\
&\geq \|D_T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\| - \|(D_T - \Lambda_T(\boldsymbol{\theta}^*))(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\| \\
&\geq (a - K\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|)\cdot\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|
\end{aligned}
\tag{19}
$$

The penultimate inequality holds by the triangle inequality, and the final inequality follows from Assumption 2(E) on the minimum eigenvalue of D_T. Thus, for T sufficiently large so that $a - K\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| > 0$, the result follows. ■

Lemma 4 *Define*

$$
\mu_t(\boldsymbol{\theta}, d) = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d}\|g_t(\boldsymbol{\tau}) - g_t(\boldsymbol{\theta})\| \tag{20}
$$

Then under assumptions 1-2, Assumption N3(ii) of Weiss (1991) holds

$$
\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)] \leq bd, \text{ for } \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| + d \leq d_0, d \geq 0 \tag{21}
$$

for T sufficiently large, where b, d,and d_0 are strictly positive numbers.

Proof of Lemma 4. In this proof, the strictly positive constant c and the mean-value expansion term, $\boldsymbol{\tau}^*$, can change from line to line. Pick d_0 such that for any $\boldsymbol{\theta}$ that satisfies

$\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \le d_0$, all the conditions in Assumption 2(C) and 2(D) hold as well as $e_t(\boldsymbol{\theta}) \le v_t(\boldsymbol{\theta}) \le 0$. Let us expand $g_t(\boldsymbol{\theta})$ into six terms:

$$\begin{aligned} g_t(\boldsymbol{\theta}) = \frac{1}{\alpha}\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\mathbf{1}\{Y_t \le v_t(\boldsymbol{\theta})\} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} + \frac{1}{\alpha}\frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2}\mathbf{1}\{Y_t \le v_t(\boldsymbol{\theta})\} \\ - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} - \frac{1}{\alpha}\frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2}\mathbf{1}\{Y_t \le v_t(\boldsymbol{\theta})\}Y_t + \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})} \end{aligned} \tag{22}$$

We will bound $\mu_t(\boldsymbol{\theta}, d)$ by considering six terms, $\mu_t(\boldsymbol{\theta}, d)^{(i)}, i = 1, 2, \cdots, 6$, defined below. Each term is shown to be bounded by a constant times d.

First term:

$$\mu_t(\boldsymbol{\theta}, d)^{(1)} = \frac{1}{\alpha}\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})}\mathbf{1}\{Y_t \le v_t(\boldsymbol{\tau})\} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\mathbf{1}\{Y_t \le v_t(\boldsymbol{\theta})\}\right\| \tag{23}$$

Set $\boldsymbol{\tau}_1 = \arg\min_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d} v_t(\boldsymbol{\tau})$ and $\boldsymbol{\tau}_2 = \arg\max_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d} v_t(\boldsymbol{\tau})$. Since $v_t(\boldsymbol{\theta})$ and $e_t(\boldsymbol{\theta})$ are assumed to be twice continously differentiable, $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ exist. We want to take the indicator function out from the 'sup' operator. To this end, let us discuss what $\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)}$ equals in two cases.

Case 1: $Y_t \le v_t(\boldsymbol{\theta})$. (a) If $Y_t > v_t(\boldsymbol{\tau}_2)$, $\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)} = \left\|\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\|$. (b) If $Y_t < v_t(\boldsymbol{\tau}_1)$, $\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)} = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\|$. (c) If $v_t(\boldsymbol{\tau}_1) \le Y_t \le v_t(\boldsymbol{\tau}_2)$,

$$\begin{aligned} \alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)} &= \max\left\{\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d, Y_t \le v(\boldsymbol{\tau})}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\|, \left\|\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\|\right\} \\ &\le \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\| + \left\|\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\| \end{aligned} \tag{24}$$

Case 2: $Y_t > v_t(\boldsymbol{\theta})$,

$$\begin{aligned} \alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)} &= \mathbf{1}\{Y_t \le v(\boldsymbol{\tau}_2)\} \cdot \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d, Y_t \le v(\boldsymbol{\tau})}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})}\right\| \\ &\le \mathbf{1}\{Y_t \le v(\boldsymbol{\tau}_2)\} \cdot \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})}\right\| \end{aligned} \tag{25}$$

$\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| + d \le d_0$ implies that both $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ (which are in a d-neighborhood of $\boldsymbol{\theta}$) are in a d_0-neighborhood of $\boldsymbol{\theta}_0$, and so

$$\left\|\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\| \le \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\le d}\left\|\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})}\right\| \le \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\le d_0}\left\|\frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})}\right\| \tag{26}$$

Thus,

$$\begin{aligned}
&\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)} \\
&\leq (\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}) \\
&\cdot \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \leq d_0} \left\| \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| + \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\|,
\end{aligned} \tag{27}$$

where

$$\begin{aligned}
\mathbb{E}_{t-1}[\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\}] &= \int_{v_t(\boldsymbol{\tau}_2)}^{v_t(\boldsymbol{\theta})} f_t(y)dy \\
&\leq K|v_t(\boldsymbol{\tau}_2) - v_t(\boldsymbol{\theta})| \leq KV_1(\mathcal{F}_{t-1})\|\boldsymbol{\tau}_2 - \boldsymbol{\theta}\| \leq KV_1(\mathcal{F}_{t-1})d
\end{aligned} \tag{28}$$

and similarly,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}|\mathcal{F}_{t-1}\right] &\leq KV_1(\mathcal{F}_{t-1})d \\
\text{and } \mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_1) < Y_t \leq v_t(\boldsymbol{\theta})\}|\mathcal{F}_{t-1}\right] &\leq KV_1(\mathcal{F}_{t-1})d
\end{aligned} \tag{29}$$

Further

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \leq HV_1(\mathcal{F}_{t-1}) \tag{30}$$

and by the mean-value theorem,

$$\frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} = \left\| \frac{\nabla^2 v_t(\boldsymbol{\tau}^*)}{-e_t(\boldsymbol{\tau}^*)} + \frac{\nabla' v_t(\boldsymbol{\tau}^*)\nabla e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^2} \right\| \cdot (\boldsymbol{\tau} - \boldsymbol{\theta}) \tag{31}$$

$$\Rightarrow \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \leq \left(HV_2(\mathcal{F}_{t-1}) + H^2 V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})\right) \cdot d. \tag{32}$$

By Assumption 2(D), $\mathbb{E}[V_2(\mathcal{F}_{t-1})]$ and $\mathbb{E}[V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})]$ are finite, so $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(1)}] \leq cd$, where c is a strictly positive constant.

Second term: $\mu_t(\boldsymbol{\theta}, d)^{(2)} = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\|$. It was shown in the derivations for the first term that $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(2)}] \leq cd$, where c is a strictly positive constant.

Third term:

$$\mu_t(\boldsymbol{\theta}, d)^{(3)} = \frac{1}{\alpha} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\tau})\} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} \right\| \tag{33}$$

Similar to the first term, $\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(3)}$ can be bounded by

$$(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}) \tag{34}$$
$$\cdot \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \leq d_0} \left\| \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| + \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\|$$

where

$$\mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\} | \mathcal{F}_{t-1}\right] \leq 3KV_1(\mathcal{F}_{t-1})d \tag{35}$$

and

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \leq d} \left\| \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \leq H \cdot H_1(\mathcal{F}_{t-1}) \tag{36}$$

where $e_t(\boldsymbol{\theta}) \leq v_t(\boldsymbol{\theta}) \leq 0$ is used, and by the mean-value theorem,

$$\frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \tag{37}$$
$$= \left\| \frac{\nabla' e_t(\boldsymbol{\tau}^*)\nabla v_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^2} - \frac{2v_t(\boldsymbol{\tau}^*)\nabla' e_t(\boldsymbol{\tau}^*)\nabla e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^3} + \frac{v_t(\boldsymbol{\tau}^*)\nabla^2 e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^2} \right\| \cdot (\boldsymbol{\tau} - \boldsymbol{\theta})$$

$$\Rightarrow \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \tag{38}$$
$$\leq \left(H^2 V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1}) + 2H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})\right) \cdot d$$

By Assumption 2(D), $\mathbb{E}[V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})], \mathbb{E}[H_1(\mathcal{F}_{t-1})^2], \mathbb{E}[H_2(\mathcal{F}_{t-1})] < \infty$. Therefore, $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(3)}] \leq cd$, where c is a strictly positive constant.

Fourth term: $\mu_t(\boldsymbol{\theta}, d)^{(4)} = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\|$. In the derivations for the third term we showed that $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(4)}] \leq cd$, where c is a strictly positive constant.

Fifth term:

$$\mu_t(\boldsymbol{\theta}, d)^{(5)} = \frac{1}{\alpha} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\tau})\} Y_t - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} Y_t \right\| \tag{39}$$

Similar to the first term, $\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(5)}$ can be bounded by

$$(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}) \tag{40}$$
$$\cdot |Y_t| \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\| \leq d_0} \left\| \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| + |Y_t| \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\|$$

where

$$\mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\}|Y_t|\ |\mathcal{F}_{t-1}\right] = \int_{v_t(\boldsymbol{\tau}_2)}^{v_t(\boldsymbol{\theta})} |y| f_t(y)dy \leq K|v_t(\boldsymbol{\tau}_2)| \cdot |v_t(\boldsymbol{\tau}_2) - v_t(\boldsymbol{\theta})| \tag{41}$$
$$\leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})\|\boldsymbol{\tau}_2 - \boldsymbol{\theta}\| \leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d$$

and similarly,

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_1) < Y_t \leq v_t(\boldsymbol{\theta})\}|Y_t|\ |\mathcal{F}_{t-1}\right] &\leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d \\ \text{and}\quad \mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}|Y_t|\ |\mathcal{F}_{t-1}\right] &\leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d \end{aligned} \tag{42}$$

Further

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \leq H^2 H_1(\mathcal{F}_{t-1}) \tag{43}$$

and by the mean-value theorem,

$$\frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} = \left\| -\frac{2\nabla' e_t(\boldsymbol{\tau}^*)\nabla e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^3} + \frac{\nabla^2 e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^2} \right\| \cdot (\boldsymbol{\tau} - \boldsymbol{\theta}) \tag{44}$$
$$\Rightarrow \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \leq \left(2H^3 H_1(\mathcal{F}_{t-1})^2 + H^2 H_2(\mathcal{F}_{t-1})\right) \cdot d$$

By Assumption 2(D), $\mathbb{E}[V(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_1(\mathcal{F}_{t-1})^2|Y_t|]$, $\mathbb{E}[H_2(\mathcal{F}_{t-1})|Y_t|] < \infty$. Therefore, $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(5)}] \leq cd$, where c is a strictly positive constant.

Sixth term:

$$\mu_t^{(6)}(\boldsymbol{\theta}, d) = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \tag{45}$$

By the mean-value theorem,

$$\frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} = \left\| \frac{\nabla' e_t(\boldsymbol{\tau}^*)\nabla e_t(\boldsymbol{\tau}^*)}{e_t(\boldsymbol{\tau}^*)^2} + \frac{\nabla^2 e_t(\boldsymbol{\tau}^*)}{-e_t(\boldsymbol{\tau}^*)} \right\| \cdot (\boldsymbol{\tau} - \boldsymbol{\theta}) \tag{46}$$

$$\Rightarrow \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \leq \left(H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})\right) \cdot d. \tag{47}$$

By Assumption 2(D), $\mathbb{E}[H_1(\mathcal{F}_{t-1})^2]$, $\mathbb{E}[H_2(\mathcal{F}_{t-1})] < \infty$. Therefore, $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(6)}] \leq cd$, where c is a strictly positive constant.

Thus we have shown that $\mu_t(\boldsymbol{\theta}, d) \leq \sum_{i=1}^{6} \mu_t(\boldsymbol{\theta}, d)^{(i)}$ with $\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^{(i)}] \leq cd$, $\forall i = 1, 2, \cdots, 6$, where c is a strictly positive constant, proving the lemma. ■

Lemma 5 *Under Assumptions 1-2, Assumption N3(iii) of Weiss (1991) holds:*

$$\mathbb{E}[\mu_t(\boldsymbol{\theta}, d)^q] \leq cd, \text{ for } \|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| + d \leq d_0, \text{ and some } q > 2$$

for T sufficiently large, and where $c > 0$, $d \geq 0$ and $d_0 > 0$.

Proof of Lemma 5. In this proof, the strictly positive constant c and the mean-value expansion term, $\boldsymbol{\tau}^*$, can change from line to line. Pick d_0 such that for any $\boldsymbol{\theta}$ that satisfies $\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \leq d_0$, all the conditions in Assumption 2(C) and 2(D) hold as well as $e_t(\boldsymbol{\theta}) \leq v_t(\boldsymbol{\theta}) \leq 0$. Similar to Lemma 4, we will decompose $\mu_t(\boldsymbol{\theta}, d)$ into six terms, $\mu_t(\boldsymbol{\theta}, d)^{(i)}$, for $i = 1, 2, ..., 6$. By Jensen's inequality, $\mathbb{E}\left[\mu_t(\boldsymbol{\theta}, d)^q\right] \leq 6^{q-1} \sum_{i=1}^{6} \mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(i)}\right)^q]$, $q > 2$. We will show that for some $0 < \delta < 1$, $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(i)}\right)^{2+\delta}] \leq cd$, $\forall\ i = 1, 2, \cdots, 6$, where c is a strictly positive constant.

First term:

$$\mu_t(\boldsymbol{\theta}, d)^{(1)} = \frac{1}{\alpha} \sup_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\tau})\} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} \right\| \tag{48}$$

Set $\boldsymbol{\tau}_1 = \arg\min_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} v_t(\boldsymbol{\tau})$ and $\boldsymbol{\tau}_2 = \arg\max_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} v_t(\boldsymbol{\tau})$. Following the same argument as in the proof of Lemma 4, we obtain

$$\begin{aligned} [\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(1)}]^{2+\delta} \leq \quad & \left(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}\right) \\ & \cdot \left(\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \leq d_0} \left\| \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{2+\delta} + \left(\sup_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{2+\delta} \end{aligned} \tag{49}$$

where

$$\mathbb{E}_{t-1}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}\right] \leq 3KV_1(\mathcal{F}_{t-1})d \tag{50}$$

and

$$\left(\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{2+\delta} \leq \left(HV_1(\mathcal{F}_{t-1})\right)^{2+\delta} \tag{51}$$

For $\left(\sup_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{2+\delta}$, we need to combine the two following two results:

$$\sup_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \leq \left(HV_2(\mathcal{F}_{t-1}) + H^2 V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1})\right) d \tag{52}$$

$$\left(\sup_{\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{1+\delta} \leq \left(2HV_1(\mathcal{F}_{t-1})\right)^{1+\delta}$$

Combining with Assumption 2(D), we thus have $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(1)}\right)^{2+\delta}] \leq cd$.

Second term: $\mu_t(\boldsymbol{\theta}, d)^{(2)} = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' v_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' v_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\|$. It was shown in the derivations for the first term that $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(2)}\right)^{2+\delta}] \leq cd$.

Third term:

$$\mu_t(\boldsymbol{\theta}, d)^{(3)} = \frac{1}{\alpha} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\tau})\} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} \right\| \tag{53}$$

Similar to the first term, $\left(\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(3)}\right)^{2+\delta}$ can be bounded by

$$\begin{aligned} &\left(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}\right) \\ &\cdot \left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d_0} \left\| \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta} + \left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta} \end{aligned} \tag{54}$$

where

$$\mathbb{E}_{t-1}\left(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}\right) \leq 3KV_1(\mathcal{F}_{t-1})d \tag{55}$$

and

$$\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d} \left\| \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta} \leq \left(H \cdot H_1(\mathcal{F}_{t-1})\right)^{2+\delta} \tag{56}$$

For $\left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta}$, we need to combine the following two results:

$$\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \leq \left(H^2 V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1}) + 2H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})\right) d \tag{57}$$

$$\left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{1+\delta} \leq \left(2H \cdot H_1(\mathcal{F}_{t-1})\right)^{1+\delta} \tag{58}$$

Combining with Assumption 2(D), we thus have $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(3)}\right)^{2+\delta}] \leq cd$.

Fourth term: $\mu_t(\boldsymbol{\theta}, d)^{(4)} = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{v_t(\boldsymbol{\tau})\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{v_t(\boldsymbol{\theta})\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\|$. It was shown in the derivations for the third term that $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta}, d)^{(4)}\right)^{2+\delta}] \leq cd$.

Fifth term:

$$\mu_t(\boldsymbol{\theta}, d)^{(5)} = \frac{1}{\alpha} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\tau})\} Y_t - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} Y_t \right\| \tag{59}$$

Similar to the first and third terms, $\left(\alpha \cdot \mu_t(\boldsymbol{\theta}, d)^{(5)}\right)^{2+\delta}$ can be bounded by

$$\begin{aligned} &(\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\tau}_1) \leq Y_t \leq v_t(\boldsymbol{\theta})\} + \mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}) \qquad (60) \\ &\cdot |Y_t|^{2+\delta} \left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d_0} \left\| \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta} + |Y_t|^{2+\delta} \left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta} \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_{t-1}[\mathbf{1}\{v_t(\boldsymbol{\tau}_2) < Y_t \leq v_t(\boldsymbol{\theta})\}|Y_t|^{2+\delta}] &= \int_{v_t(\boldsymbol{\tau}_2)}^{v_t(\boldsymbol{\theta})} |y|^{2+\delta} f_t(y) dy \leq K|v_t(\boldsymbol{\tau}_2)|^{2+\delta} \cdot |v_t(\boldsymbol{\tau}_2) - v_t(\boldsymbol{\theta})| \qquad (61) \\ &\leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) \|\boldsymbol{\tau}_2 - \boldsymbol{\theta}\| \leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \end{aligned}$$

and similarly,

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\tau}_1) < Y_t \leq v_t(\boldsymbol{\theta})\}|Y_t|^{2+\delta} \,|\mathcal{F}_{t-1}\right] &\leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \qquad (62) \\ \text{and } \mathbb{E}\left[\mathbf{1}\{v_t(\boldsymbol{\theta}) < Y_t \leq v_t(\boldsymbol{\tau}_2)\}|Y_t|^{2+\delta} \,|\mathcal{F}_{t-1}\right] &\leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \end{aligned}$$

Further

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \leq H^2 H_1(\mathcal{F}_{t-1}) \tag{63}$$

For $\left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{2+\delta}$, we need to combine the following two results:

$$\begin{aligned} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{e_t(\boldsymbol{\tau})^2} - \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| &\leq \left(2H^3 H_1(\mathcal{F}_{t-1})^2 + H^2 H_2(\mathcal{F}_{t-1})\right) d \qquad (64) \\ \left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})^2} \right\| \right)^{1+\delta} &\leq \left(2H^2 H_1(\mathcal{F}_{t-1})\right)^{1+\delta} \end{aligned}$$

Combining with Assumption 2(D), we thus have $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta} d)^{(5)}\right)^{2+\delta}] \leq cd$.

Sixth term:

$$\mu_t^{(6)}(\boldsymbol{\theta}, d) = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \tag{65}$$

We have

$$\begin{aligned} \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| &\leq \left(H^2 H_1(\mathcal{F}_{t-1})^2 + H H_2(\mathcal{F}_{t-1})\right) d \qquad (66) \\ \left(\sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\|\leq d} \left\| \frac{\nabla' e_t(\boldsymbol{\tau})}{-e_t(\boldsymbol{\tau})} - \frac{\nabla' e_t(\boldsymbol{\theta})}{-e_t(\boldsymbol{\theta})} \right\| \right)^{1+\delta} &\leq \left(2H H_1(\mathcal{F}_{t-1})\right)^{1+\delta} \end{aligned}$$

Combining with Assumption 2(D), we thus have $\mathbb{E}[\left(\mu_t(\boldsymbol{\theta},d)^{(6)}\right)^{2+\delta}] \leq cd$. Thus $\mathbb{E}[\mu_t(\boldsymbol{\theta},d)^{(i)}]^{2+\delta} \leq cd$, $\forall i = 1,2,...,6$, proving the lemma. ■

Lemma 6 *Under Assumptions 1-2, $E\|g_t(\boldsymbol{\theta}^0)\|^{2+\delta} \leq M$, for all t and some $M > 0$.*

Proof of Lemma 6.

$$\begin{aligned}
\mathbb{E}\|g_t(\boldsymbol{\theta}^0)\|^{2+\delta} \leq & 4^{1+\delta}\left\{\mathbb{E}\left\|\frac{\nabla' v_t(\boldsymbol{\theta}^0)}{-e_t(\boldsymbol{\theta}^0)}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t(\boldsymbol{\theta}^0)\right\}-1\right)\right\|^{2+\delta}\right. \\
& + \mathbb{E}\left\|\frac{v_t(\boldsymbol{\theta}^0)\nabla' e_t(\boldsymbol{\theta}^0)}{e_t(\boldsymbol{\theta}^0)^2}\left(\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t(\boldsymbol{\theta}^0)\right\}-1\right)\right\|^{2+\delta} \\
& + \mathbb{E}\left\|\frac{\nabla' e_t(\boldsymbol{\theta}^0)}{e_t(\boldsymbol{\theta}^0)^2}\frac{1}{\alpha}\mathbf{1}\left\{Y_t \leq v_t(\boldsymbol{\theta}^0)\right\}Y_t\right\|^{2+\delta} \\
& \left. + \mathbb{E}\left\|\frac{\nabla' e_t(\boldsymbol{\theta}^0)}{e_t(\boldsymbol{\theta}^0)}\right\|^{2+\delta}\right\} \\
\leq & 4^{1+\delta}\left\{(\frac{1}{\alpha}+1)^{2+\delta}H^{2+\delta}\mathbb{E}\left[V_1(\mathcal{F}_{t-1})^{2+\delta}\right]\right. \\
& + (\frac{1}{\alpha}+1)^{2+\delta}H^{2+\delta}\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{2+\delta}\right] \\
& + \frac{1}{\alpha^{2+\delta}}H^{4+2\delta}\mathbb{E}[H_1(\mathcal{F}_{t-1})^{2+\delta}\,|Y_t|^{2+\delta}] \\
& \left. + H^{2+\delta}\mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{2+\delta}\right]\right\} \\
\leq & M
\end{aligned}$$

since all the four expectations in the penultimate inequality are finite by assumption 2(D). Assumption N4 of Weiss (1991) only requires $E\|g_t(\boldsymbol{\theta}^0)\|^2 \leq M$, which is implied by the above. ■

Lemma 7 *Under Assumptions 1-2, we have $T^{-1/2}\sum_{t=1}^{T} g_t(\boldsymbol{\theta}^{\mathbf{0}}) \xrightarrow{d} N(0,\mathbf{A}_0)$ as $T \to \infty$, where $\mathbf{A}_0 \equiv \mathbb{E}\left[g_t(\boldsymbol{\theta}^{\mathbf{0}})g_t(\boldsymbol{\theta}^{\mathbf{0}})'\right]$.*

Proof of Lemma 7. First note that the sequence $\left\{g_t(\boldsymbol{\theta}^{\mathbf{0}})\right\}$ is stationary by Assumption 1(B)(ii), and has zero mean. Under Assumption 2(F) and Lemma 6, we can use Corollary 5.1 of Hall and Heyde (1980) and the Cramer-Wold device to obtain the result. ■

Appendix SA.2: Estimating a GARCH(1,1) model by FZ loss minimization

In this appendix we show that we can estimate the popular GARCH(1,1) model via FZ loss minimization. We then verify that the assumptions required to show this are implied by the Assumptions 1-2 in the main paper. Throughout, $||\mathbf{x}||$ refers to the Euclidean norm if $\mathbf{x}$ is a vector and to the Frobenius norm if $\mathbf{x}$ is a matrix.

Appendix SA.2.1: Model specification

Assume that the data generating process for Y_t is:

$$\begin{aligned} Y_t &= \sigma_t \eta_t, \quad \eta_t \perp\!\!\!\perp \sigma_t, \quad \eta_t \sim iid\ F_\eta(0,1) \\ \sigma_t^2 &= \omega_0 + \beta_0 \sigma_{t-1}^2 + \gamma_0 Y_{t-1}^2 \end{aligned} \tag{67}$$

Under this model, the conditional VaR and ES of Y_t at a probability level $\alpha \in (0,1)$, that is $VaR_\alpha(Y_t|\mathcal{F}_{t-1})$ and $ES_\alpha(Y_t|\mathcal{F}_{t-1})$, follow the dynamics:

$$\begin{aligned} \begin{bmatrix} VaR_\alpha(Y_t|\mathcal{F}_{t-1}) \\ ES_\alpha(Y_t|\mathcal{F}_{t-1}) \end{bmatrix} &= \begin{bmatrix} c_0 \cdot ES_\alpha(Y_t|\mathcal{F}_{t-1}) \\ b_0 \cdot \sigma_t \end{bmatrix} \\ \text{where } c_0 &\equiv F_\eta^{-1}(\alpha)/\mathbb{E}[\eta_t|\eta_t \le F_\eta^{-1}(\alpha)] \\ b_0 &\equiv \mathbb{E}[\eta_t|\eta_t \le F_\eta^{-1}(\alpha)] \end{aligned} \tag{68}$$

We fix the level $\alpha \in (0,1)$ throughout this appendix. Our goal is to estimate the parameter vector $\boldsymbol{\theta}^0 = [\beta_0, \gamma_0, b_0, c_0]$ by minimizing the FZ loss function. Note that the parameters do not include ω_0 because only two of the three parameters ω_0, b_0, γ_0 are identifiable under this model. A detailed discussion about the identification of the GARCH model via FZ loss minimization is provided in Section SA.2.3 of this appendix.

In the simulation study (Section 4 of the main paper), for estimating the GARCH model via FZ loss minimization, we fix ω at its true value ω_0. Put $\boldsymbol{\theta} = [\beta, \gamma, b, c]$ and its true value is

$\boldsymbol{\theta}^0 = [\beta_0, \gamma_0, b_0, c_0]$. We will estimate $\boldsymbol{\theta}^0$ by

$$\begin{aligned} \boldsymbol{\theta}_T &\equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \; L_T(\boldsymbol{\theta}) \\ \text{where} \;\; L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T} L_{FZ0}(Y_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}); \alpha) \\ \sigma_t^2(\boldsymbol{\theta}) &= \omega_0 + \beta\sigma_{t-1}^2(\boldsymbol{\theta}) + \gamma Y_{t-1}^2 \\ v_t(\boldsymbol{\theta}) &= c \cdot e_t(\boldsymbol{\theta}) \\ e_t(\boldsymbol{\theta}) &= b \cdot \sigma_t(\boldsymbol{\theta}) \end{aligned}$$

and the FZ loss function L_{FZ0} is defined in equation (6)

Appendix SA.2.2: Assumptions to estimate GARCH by FZ minimization

GARCH Assumption 1: F_η has zero mean, unit variance, finite fourth moment, and a unique α-quantile, which is non-positive. It has density $f_\eta(\cdot)$ that satisfies $f_\eta(\cdot) \leq K$ and $|f_\eta(\lambda_1) - f_\eta(\lambda_2)| \leq K|\lambda_1 - \lambda_2|$.

The distributions we often assume for the innovations of GARCH model, like the normal distribution or t-distribution with degrees of freedom greater than four, all satisfy this assumption.

GARCH Assumption 2: $0 < \omega_0 < \infty$. The true parameter vector $\boldsymbol{\theta}^0 = [\beta_0, \gamma_0, b_0, c_0] \in \Theta \in \mathbb{R}^4$ is in the interior of Θ, a compact and convex parameter space. Specifically, for any vector $[\beta, \gamma, b, c] \in \Theta$, assume that $\delta_1 \leq \beta \leq (1-\delta_1)$, $\delta_1 \leq \gamma \leq (1-\delta_1)$ for some constant $\delta_1 > 0$, $\delta_2 \leq c \leq (1-\delta_2)$, $-B_1 \leq b \leq -B_2$, for some constants $\delta_2, B_1, B_2 > 0$, and $(\beta+\gamma)^2 + \left(\mathbb{E}\left[\eta_t^4\right] - 1\right)\gamma^2 \leq 1 - \delta_3$ for some constant $\delta_3 > 0$.

This assumption is similar to Assumption 1 of Lumsdaine (1996) with the exception of the third condition on the parameter vector, which is used to validate the mixing condition in Assumption 2(F), which we now discuss. It is not hard to show that

$$\sigma\left\{\left(Y_t, v_t\left(\theta\right), e_t\left(\theta\right), \nabla' v_t\left(\theta\right), \nabla' e_t\left(\theta\right)\right)\right\} \subset \sigma\left\{\left(Y_t, \sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right)\right\}$$

and thus we need to consider the mixing properties of $\left(Y_t, \sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right)$. Using Definition 3 of Carrasco and Chen (2002), $\left\{\left(Y_t, \sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right), t \geq 0\right\}$ is a generalized hidden Markov model with a hidden chain $\left\{\left(\sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right), t \geq 0\right\}$. By their Proposition 4, if $\left(\sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right)$ is stationary and β-mixing then $\left(Y_t, \sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right)$ is stationary and β-mixing with a decay rate at least as fast as that of $\left\{\left(\sigma_t^2\left(\theta\right), \partial\sigma_t^2\left(\theta\right)/\beta\right), t \geq 0\right\}$.

We use Proposition 3 of Carrasco and Chen (2002). First, we express $\left\{\left(\sigma_t^2(\theta), \partial\sigma_t^2(\theta)/\beta\right), t \geq 0\right\}$ in the polynomial random coefficient form:

$$\begin{pmatrix} \sigma_t^2(\theta) \\ \partial\sigma_t^2(\theta)/\beta \end{pmatrix} = \begin{pmatrix} \omega_0 \\ 0 \end{pmatrix} + \begin{pmatrix} \gamma\eta_{t-1}^2 + \beta & 0 \\ 1 & \beta \end{pmatrix} \begin{pmatrix} \sigma_{t-1}^2(\theta) \\ \partial\sigma_{t-1}^2(\theta)/\beta \end{pmatrix} \tag{69}$$

First, note that by GARCH Assumption 1 $\left\{\eta_t^2\right\}$ satisfies their condition (e) and that by GARCH Assumption 2, their Assumption A_0 is obviously satisfied. For Assumption A_1, the spectral radius of $\begin{pmatrix} \beta & 0 \\ 1 & \beta \end{pmatrix}$ is $\beta < 1$. For Assumption A_2', the spectral radius of $\begin{pmatrix} \gamma\eta_{t-1}^2 + \beta & 0 \\ 1 & \beta \end{pmatrix}$ is $\mathbb{E}\left[\left(\gamma\eta_{t-1}^2 + \beta\right)^2\right] = (\beta+\gamma)^2 + \left(\mathbb{E}\left[\eta_t^4\right] - 1\right)\gamma^2 < 1$.Then, if $\left(\sigma_t^2(\theta), \partial\sigma_t^2(\theta)/\beta\right)$ is initialized from the invariant distribution (which we did in our simulations) then $\left\{\left(\sigma_t^2(\theta), \partial\sigma_t^2(\theta)/\beta\right), t \geq 0\right\}$ is strictly stationary and β-mixing with exponential decay. It is well known that β-mixing implies α-mixing and so Assumption 2(F) of the paper is satisfied.

GARCH Assumptions 1–2 imply that the distribution and density of Y_t conditional on $\mathcal{F}_{t-1}$ satisfy Assumption 2(B)(i). Since $Y_t = \sigma_t\eta_t$ and $\sigma_t \in \mathcal{F}_{t-1}$,

$$\begin{aligned} F_t(x|\mathcal{F}_{t-1}) =& F_\eta(\frac{x}{\sigma_t}) \\ f_t(x|\mathcal{F}_{t-1}) =& \frac{1}{\sigma_t} f_\eta(\frac{x}{\sigma_t}) \end{aligned}$$

Thus,

$$|f_t(x|\mathcal{F}_{t-1})| \leq \frac{K}{\sqrt{\omega_0}}, \text{ since } \sigma_t^2 = \omega_0 + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 \geq \omega_0 > 0$$

$$|f_t(\lambda_1|\mathcal{F}_{t-1}) - f_t(\lambda_2|\mathcal{F}_{t-1})| = \frac{1}{\sigma_t}|f_\eta(\frac{\lambda_1}{\sigma_t}) - f_\eta(\frac{\lambda_1}{\sigma_t})| \leq \frac{K}{\sigma_t^2}|\lambda_1 - \lambda_2| \leq \frac{K}{\omega_0}|\lambda_1 - \lambda_2|$$

GARCH Assumption 3: $\mathbb{E}|Y_t|^{5+\delta} < \infty$, for some $\delta > 0$.

GARCH Assumption 3 is needed to show the uniform LLN Assumption 1(A) of the paper and also to ensure the moment conditions in Assumptions 2 (C) and (D).

For the GARCH model it is possible to obtain the results of the paper under a weaker version of Assumption 2(D). An inspection of the proofs shows that it is sufficient to replace Assumption 2(D) by the following.

Assumption 2(D'): For some $0 < \delta < 1$ and $\forall\ t$:
(i) $\mathbb{E}\left[V_1(\mathcal{F}_{t-1})^{3+\delta}\right], \mathbb{E}\left[H_1(\mathcal{F}_{t-1})^{3+\delta}\right], \mathbb{E}\left[V_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}}\right], \mathbb{E}\left[H_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}}\right] \leq K$,

(ii) $\mathbb{E}\left[V\left(\mathcal{F}_{t-1}\right)^{2+\delta} V_1\left(\mathcal{F}_{t-1}\right)^{1+\delta}\right] \leq K$,
(iii) $\mathbb{E}\left[H_1\left(\mathcal{F}_{t-1}\right)^{2+\delta}\left|Y_t\right|^{2+\delta}\right], \mathbb{E}\left[H_2\left(\mathcal{F}_{t-1}\right)^{1+\delta}\left|Y_t\right|^{2+\delta}\right] \leq K$.

Assumption 2(D') is in turn fulfilled if $\mathbb{E}|Y_t|^{4+\delta} < \infty$, for some $\delta > 0$. For reasons of brevity, we omit the arguments and work instead with the stronger GARCH Assumption 3.

Appendix SA.2.3: Identification

In Theorem 1, we discussed the identification of a general dynamic model for ES and VaR model by minimizing the FZ loss, with the form of a general model given by equation (4). Under correct specification of the model, that is $(VaR_\alpha(Y_t|\mathcal{F}_{t-1}), ES_\alpha(Y_t|\mathcal{F}_{t-1})) = (v_t(\boldsymbol{\theta}^0), e_t(\boldsymbol{\theta}^0))\ \forall\, t$ a.s., the condition required for identification is given by Assumption 1(B) (iv): $\Pr\left[v_t(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}^0) \cap e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0)\right] = 1\ , \forall\ t \Rightarrow \theta = \theta^0$. This assumption is equivalent to

$$\Pr\left[v_t(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}^0) \cap e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0),\ \forall t\right] = 1$$

In the case of the GARCH model we have:

$$\begin{aligned}
&\Pr\left[\{v_t(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}^0)\} \cap \{e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0)\},\ \ \forall t\right] = 1 \\
\Rightarrow &\Pr\left[\{c \cdot e_t(\boldsymbol{\theta}) = c_0 \cdot e_t(\boldsymbol{\theta}^0)\} \cap \{e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0)\},\ \ \forall t\right] = 1 \\
\Rightarrow &\Pr\left[\{c = c_0\} \cap \{b \cdot \sigma_t(\boldsymbol{\theta}) = b_0 \cdot \sigma_t(\boldsymbol{\theta}^0)\},\ \ \forall t\right] = 1 \\
\Rightarrow &c = c_0, \quad \Pr\left[b^2 \cdot \sigma_t^2(\boldsymbol{\theta}) = b_0^2 \cdot \sigma_t^2(\boldsymbol{\theta}^0),\ \ \forall t\right] = 1 \\
\Rightarrow &c = c_0, \quad \Pr\left[b^2(\omega + \beta\sigma_{t-1}^2(\boldsymbol{\theta}) + \gamma Y_{t-1}^2) = b_0^2(\omega_0 + \beta_0\sigma_{t-1}^2(\boldsymbol{\theta}^0) + \gamma_0 Y_{t-1}^2),\ \ \forall t\right] = 1 \\
\Rightarrow &c = c_0, \quad \Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) + b^2\gamma Y_{t-1}^2 = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) + b_0^2\gamma_0 Y_{t-1}^2,\ \ \forall t\right] = 1
\end{aligned}$$

where the third line holds because $e_t(\boldsymbol{\theta}^0) = b_0\sigma_t(\boldsymbol{\theta}^0)$ and we assume that $b_0 < 0$, thus $e_t(\boldsymbol{\theta}^0) < 0$, and in the last line, we replaced $b^2\sigma_{t-1}^2(\boldsymbol{\theta})$ by $b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0)$ because we started with $b^2\sigma_t^2(\boldsymbol{\theta}) = b_0^2\sigma_t^2(\boldsymbol{\theta}^0), \forall t$ almost surely.

Since the GARCH model assumes that $Y_{t-1}|\sigma_{t-1}(\boldsymbol{\theta}^0) \sim F_\eta(0, \sigma_{t-1}^2(\boldsymbol{\theta}^0))$,

$$\begin{aligned}
&\Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) + b^2\gamma Y_{t-1}^2 = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) + b_0^2\gamma_0 Y_{t-1}^2,\ \forall t\right] = 1 \\
\Rightarrow &\Pr\left[\{b^2\gamma = b_0^2\gamma_0\} \cap \{b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0)\},\ \forall t\right] = 1 \\
\Rightarrow &b^2\gamma = b_0^2\gamma_0, \quad \Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0),\ \forall t\right] = 1
\end{aligned}$$

If $\beta b_0^2 \neq \beta_0 b_0^2$ and $Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0),\ \forall t\right] = 1$ hold at the same time,

then we have

$$\Pr\left[\sigma_{t-1}^2(\boldsymbol{\theta}^0) = \frac{b_0^2\omega_0 - b^2\omega}{\beta b_0^2 - \beta_0 b_0^2},\ \forall t\right] = 1$$
$$\Rightarrow \Pr\left[\omega_0 + \beta_0\sigma_{t-2}^2(\boldsymbol{\theta}^0) + \gamma_0 Y_{t-2}^2 = \frac{b_0^2\omega_0 - b^2\omega}{\beta b_0^2 - \beta_0 b_0^2},\ \forall t\right] = 1$$

This contradicts the assumption of the GARCH model,that $Y_{t-2}|\sigma_{t-2}(\boldsymbol{\theta}^0) \sim F_\eta(0, \sigma_{t-2}^2(\boldsymbol{\theta}^0))$. Thus, $\beta b_0^2 \neq \beta_0 b_0^2$ and $\Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0),\ \forall t\right] = 1$ cannot hold at the same time. This means that $\Pr\left[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\boldsymbol{\theta}^0),\ \forall t\right] = 1$ implies $\beta b_0^2 = \beta_0 b_0^2$, which further implies that $\beta = \beta_0$ and $b^2\omega = b_0^2\omega_0$. In summary, we have shown that

$$\Pr\left[v_t(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}^0) \cap e_t(\boldsymbol{\theta}) = e_t(\boldsymbol{\theta}^0)\right] = 1\ ,\forall\ t$$
$$\Rightarrow c = c_0,\ b^2\gamma = b_0^2\gamma_0,\ \beta b_0^2 = \beta_0 b_0^2,\ b^2\omega = b_0^2\omega_0$$
$$\Rightarrow c = c_0,\ \beta = \beta_0,\ b^2\gamma = b_0^2\gamma_0,\ b^2\omega = b_0^2\omega_0$$

Therefore, Assumption 1(B)(iv) holds if we normalize one of the three parameters b, γ, ω. We choose to normalize ω.

Appendix SA.2.4: Uniform LLN

In this section, we show that under the GARCH assumptions we have made in Section SA.2.2, Assumption 1(A) is satisfied: $L_{FZ0}\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right); \alpha\right)$ obeys the uniform law of large numbers.

Since the parameter space is assumed to be compact, we can establish the uniform LLN by combining the pointwise LLN with stochastic equicontinuity.

Appendix SA.2.4.1: LLN

The LLN is based on Davidson (1994, Corollary 19.3) which we restate here as Theorem 4 for convenience.

Theorem 4 (Davidson) *Suppose that* $(X_t)_{t\in\mathbb{N}}$ *satisfies:*

$$\sup_{t\in\mathbb{N}} \mathbb{E}|X_t|^{2+\delta} < \infty \text{ for some } \delta > 0$$

and $(X_t)_{t\in\mathbb{N}}$ *is* α *mixing with* $\sum_{m=1}^{\infty} m^{-1}\alpha\left(m\right)^{\delta/(2+\delta)} < \infty$, *then*

$$\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbb{E}[X_t]) \xrightarrow{L_2} 0.$$

Under Assumption 2(F), which we discussed in the context of the GARCH model in Section SA.2.2 above, implies that $L_{FZ0}\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right)\right)$ is α-mixing with a decay rate no slower than that required by Theorem 4. Also, $L_{FZ0}\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right)\right)$ is strictly stationary as we have shown that $\left(Y_t, \sigma_t^2\left(\boldsymbol{\theta}\right), \partial\sigma_t^2\left(\boldsymbol{\theta}\right)/\partial\beta\right)$ is strictly stationary. We then need only show that

$$\mathbb{E}|L_{FZ0}(Y_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}); \alpha)|^{2+\delta} < \infty$$

$$\begin{aligned}
&|L_{FZ0}(Y_t, v_t(\boldsymbol{\theta}), e_t(\boldsymbol{\theta}); \alpha)| \\
&\quad = |-\frac{1}{\alpha e_t(\boldsymbol{\theta})}\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\}(v_t(\boldsymbol{\theta}) - Y_t) + \frac{v_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})} + \log(-e_t(\boldsymbol{\theta})) - 1| \\
&\quad = |\frac{v_t(\boldsymbol{\theta})}{e_t(\boldsymbol{\theta})} \cdot (1 - \frac{1}{\alpha}\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\}) + \frac{Y_t}{\alpha e_t(\boldsymbol{\theta})}\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} + \log(-e_t(\boldsymbol{\theta})) - 1| \\
&\quad = |c \cdot (1 - \frac{1}{\alpha}\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\}) - \frac{\eta_t}{\alpha b}\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\} + \log(-b\sigma_t(\boldsymbol{\theta})) - 1| \\
&\quad \leq c(1 + \frac{1}{\alpha}) + |\log(-b)| + 1 + \frac{|\eta_t|}{\alpha|b|} + |\log\sigma_t|
\end{aligned}$$

By Cr-inequality, it is sufficient to show that

$$\mathbb{E}|\eta_t|^{2+\delta} < \infty \text{ and } \mathbb{E}|\log\sigma_t|^{2+\delta} < \infty.$$

The moment condition on η_t is directly implied by the structure of the model and GARCH Assumption 3. Recall that $\sigma_t^2 = \omega_0 + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 \geq \omega_0 > 0$. Therefore, if $\sigma_t^2 < 1$ then $|\log\sigma_t| \leq |\log\sqrt{\omega_0}|$, and if $\sigma_t^2 \geq 1$ then $|\log\sigma_t| \leq \sigma_t$. In summary, $|\log\sigma_t| \leq |\log\sqrt{\omega_0}| + \sigma_t$. Therefore, by Cr-inequality

$$\begin{aligned}
\mathbb{E}|\log\sigma_t|^{2+\delta} &\leq \mathbb{E}(|\log\sqrt{\omega_0}| + \sigma_t)^{2+\delta} \\
&\leq 2^{1+\delta}(|\log\sqrt{\omega_0}|^{2+\delta} + \mathbb{E}\sigma_t^{2+\delta}).
\end{aligned}$$

Thus, a sufficient condition for $\mathbb{E}|\log\sigma_t|^{2+\delta} < \infty$ is $\mathbb{E}\left[\sigma_t^{2+\delta}\right] < \infty$ which is implied by GARCH Assumption 3. Hence, $L_{FZ0}\left(Y_t, v_t\left(\boldsymbol{\theta}\right), e_t\left(\boldsymbol{\theta}\right)\right)$ obeys the law of large numbers for any fixed $\boldsymbol{\theta}$ by Theorem 4.

Appendix SA.2.4.2: Stochastic equicontinuity

The stochastic equicontinuity condition is derived using Davidson (1994, Theorem 21.10) which we restate here as Theorem 5 for convenience.

Theorem 5 (Davidson) *Let $Q_n(\cdot)$ be the objective function for an M-estimator. Suppose there exists $N \in \mathbb{N}$ such that*

$$|Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}')| \leq a_n h(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|), \ a.s.$$

holds for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $n \geq \mathbb{N}$, where h is a deterministic function with $h(x) \downarrow 0$ as $x \downarrow 0$, and $a_n = \mathcal{O}_p(1)$. Then $(Q_n)_{n\in\mathbb{N}}$ is stochastically equicontinuous.

Observe that

$$\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta})\}(v_t(\boldsymbol{\theta}) - Y_t) = \frac{1}{2}(v_t(\boldsymbol{\theta}) - Y_t + |v_t(\boldsymbol{\theta}) - Y_t|). \tag{70}$$

Let $\boldsymbol{\theta}_1 = [\beta_1, \gamma_1, b_1, c_1], \boldsymbol{\theta}_2 = [\beta_2, \gamma_2, b_2, c_2] \in \Theta$. Then, using (70), we obtain

$$\begin{aligned}
&\left|\frac{\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta}_1)\}(v_t(\boldsymbol{\theta}_1) - Y_t)}{e_t(\boldsymbol{\theta}_1)} - \frac{\mathbf{1}\{Y_t \leq v_t(\boldsymbol{\theta}_2)\}(v_t(\boldsymbol{\theta}_2) - Y_t)}{e_t(\boldsymbol{\theta}_2)}\right| \\
=&\frac{1}{2}\left|\frac{v_t(\boldsymbol{\theta}_1) - Y_t + |v_t(\boldsymbol{\theta}_1) - Y_t|}{e_t(\boldsymbol{\theta}_1)} - \frac{v_t(\boldsymbol{\theta}_2) - Y_t + |v_t(\boldsymbol{\theta}_2) - Y_t|}{e_t(\boldsymbol{\theta}_2)}\right| \\
\leq&\frac{1}{2}\left|\frac{v_t(\boldsymbol{\theta}_1) - Y_t}{e_t(\boldsymbol{\theta}_1)} - \frac{v_t(\boldsymbol{\theta}_2) - Y_t}{e_t(\boldsymbol{\theta}_2)}\right| + \frac{1}{2}\left|\frac{|v_t(\boldsymbol{\theta}_1) - Y_t|}{e_t(\boldsymbol{\theta}_1)} - \frac{|v_t(\boldsymbol{\theta}_2) - Y_t|}{e_t(\boldsymbol{\theta}_2)}\right| \qquad (71) \\
\leq&\left|\frac{v_t(\boldsymbol{\theta}_1) - Y_t}{e_t(\boldsymbol{\theta}_1)} - \frac{v_t(\boldsymbol{\theta}_2) - Y_t}{e_t(\boldsymbol{\theta}_2)}\right| \qquad (72) \\
=&\left|\frac{v_t(\boldsymbol{\theta}_1)}{e_t(\boldsymbol{\theta}_1)} - \frac{v_t(\boldsymbol{\theta}_2)}{e_t(\boldsymbol{\theta}_2)} - \left(\frac{Y_t}{e_t(\boldsymbol{\theta}_1)} - \frac{Y_t}{e_t(\boldsymbol{\theta}_2)}\right)\right| \\
=&\left|c_1 - c_2 - \left(\frac{\eta_t}{b_1} - \frac{\eta_t}{b_2}\right)\right| \\
\leq&|c_1 - c_2| + \frac{|b_1 - b_2|}{|b_1 b_2|}|\eta_t|.
\end{aligned}$$

The inequality between (71) and (72) holds because

$$\begin{aligned}
\left|\frac{|v_t(\boldsymbol{\theta}_1) - Y_t|}{e_t(\boldsymbol{\theta}_1)} - \frac{|v_t(\boldsymbol{\theta}_2) - Y_t|}{e_t(\boldsymbol{\theta}_2)}\right| &= \left|\frac{|v_t(\boldsymbol{\theta}_1) - Y_t|}{-|e_t(\boldsymbol{\theta}_1)|} - \frac{|v_t(\boldsymbol{\theta}_2) - Y_t|}{-|e_t(\boldsymbol{\theta}_2)|}\right| \\
&= \left|\frac{|v_t(\boldsymbol{\theta}_2) - Y_t|}{|e_t(\boldsymbol{\theta}_2)|} - \frac{|v_t(\boldsymbol{\theta}_1) - Y_t|}{|e_t(\boldsymbol{\theta}_1)|}\right| \\
&\leq \left|\frac{v_t(\boldsymbol{\theta}_2) - Y_t}{e_t(\boldsymbol{\theta}_2)} - \frac{v_t(\boldsymbol{\theta}_1) - Y_t}{e_t(\boldsymbol{\theta}_1)}\right|.
\end{aligned}$$

By Taylor's theorem,

$$|\log(-e_t(\boldsymbol{\theta}_1)) - \log(-e_t(\boldsymbol{\theta}_2))| = \left|\frac{1}{-e_t(\boldsymbol{\theta}_1^*)}\right| \cdot \|\nabla e_t(\boldsymbol{\theta}_1^*)\| \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for some $\boldsymbol{\theta}_1^*$ between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Since,

$$\begin{aligned}\|\nabla e_t(\boldsymbol{\theta})\| =& \|b \cdot \nabla \sigma_t(\boldsymbol{\theta}) + \sigma_t(\boldsymbol{\theta}) \cdot [0,0,1,0]\| \\ \leq& |b| \cdot \|\nabla \sigma_t(\boldsymbol{\theta})\| + \sigma_t(\boldsymbol{\theta}) \\ \Rightarrow \frac{\|\nabla e_t(\boldsymbol{\theta})\|}{|e_t(\boldsymbol{\theta})|} \leq& \frac{|b| \cdot \|\nabla \sigma_t(\boldsymbol{\theta})\| + \sigma_t(\boldsymbol{\theta})}{|b \cdot \sigma_t(\boldsymbol{\theta})|} \leq \frac{\|\nabla \sigma_t(\boldsymbol{\theta})\|}{\sigma_t(\boldsymbol{\theta})} + \frac{1}{|b|},\end{aligned}$$

we obtain

$$|\log(-e_t(\boldsymbol{\theta}_1)) - \log(-e_t(\boldsymbol{\theta}_2))| \leq \left(\frac{\|\nabla \sigma_t(\boldsymbol{\theta}_1^*)\|}{\sigma_t(\boldsymbol{\theta}_1^*)} + \frac{1}{|b_1^*|} \right) \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Therefore,

$$\begin{aligned}&|L_{FZ0}(Y_t, v_t(\boldsymbol{\theta}_1), e_t(\boldsymbol{\theta}_1); \alpha) - L_{FZ0}(Y_t, v_t(\boldsymbol{\theta}_2), e_t(\boldsymbol{\theta}_2); \alpha)| \\ &\quad \leq \frac{1}{\alpha} \left(|c_1 - c_2| + \frac{|b_1 - b_2|}{|b_1 b_2|} |\eta_t| \right) + |c_1 - c_2| + \left(\frac{\|\nabla \sigma_t(\boldsymbol{\theta}_1^*)\|}{\sigma_t(\boldsymbol{\theta}_1^*)} + \frac{1}{|b_1^*|} \right) \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\ &\quad \leq \left(1 + \frac{1}{\alpha} + \frac{1}{B_2} + \frac{|\eta_t|}{\alpha B_2^2} + \frac{\|\nabla \sigma_t(\boldsymbol{\theta}_1^*)\|}{\sigma_t(\boldsymbol{\theta}_1^*)} \right) \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|\end{aligned}$$

The last inequality holds because by GARCH Assumption 2, $|b_1|, |b_2|, |b_1^*| \geq B_2 > 0$.

Using Lemma 8 in Section SA.2.5, we obtain

$$\begin{aligned}\frac{\|\nabla \sigma_t(\boldsymbol{\theta})\|}{\sigma_t(\boldsymbol{\theta})} \leq& \frac{1}{2} \cdot \left[\gamma^{1/2} \beta^{-1/2} \sigma_t(\boldsymbol{\theta})^{-1} \sum_{i=2}^{\infty} (i-1) \beta^{(i-2)/2} |Y_{t-i}| + \gamma^{-1} \right] \\ \leq& \frac{1}{2} \cdot \left[(1-\delta_1)^{1/2} \delta_1^{-1/2} \omega_0^{-1/2} \sum_{i=2}^{\infty} (i-1)(1-\delta_1)^{(i-2)/2} |Y_{t-i}| + \delta_1^{-1} \right]\end{aligned}$$

using the bounds on γ and β in GARCH Assumption 2. Define

$$Z_t = \frac{1}{2} \cdot \left[(1-\delta_1)^{1/2} \delta_1^{-1/2} \omega_0^{-1/2} \sum_{i=2}^{\infty} (i-1)(1-\delta_1)^{(i-2)/2} |Y_{t-i}| + \delta_1^{-1} \right]$$

Then,

$$|L_T(\boldsymbol{\theta}_1) - L_T(\boldsymbol{\theta}_2)| \leq \left(1 + \frac{1}{\alpha} + \frac{1}{B_2} + \frac{1}{\alpha B_2^2} \cdot \frac{1}{T} \sum_{t=1}^{T} |\eta_t| + \frac{1}{T} \sum_{t=1}^{T} Z_t \right) \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

$\mathbb{E}|\eta_t| = \text{const} < \infty$ and $\mathbb{E}[Z_t] = \text{const} < \infty$ (because $\mathbb{E}|Y_t| = \text{const} < \infty$ by GARCH Assumptions 2 and 3). Then, $\frac{1}{T} \sum_{t=1}^{T} |\eta_t| = \mathcal{O}_p(1)$ and $\frac{1}{T} \sum_{t=1}^{T} Z_t = \mathcal{O}_p(1)$.

Therefore, by Theorem 5, L_T is stochastically equicontinuous.

Appendix SA.2.5: Assumptions 2(C) and 2(D)

We define

$$X_1(t;\beta) = \sum_{i=2}^{\infty}(i-1)\beta^{i-2}Y_{t-i}^2$$
$$X_2(t;\beta) = \sum_{i=2}^{\infty}(i-1)\beta^{(i-2)/2}|Y_{t-i}|$$
$$X_3(t;\beta) = \sum_{i=3}^{\infty}\frac{(i-1)(i-2)}{2}\beta^{i-3}Y_{t-i}^2$$

Lemma 8 *Under GARCH Assumption 2, we have*

$$\begin{aligned}
\|\nabla\sigma_t^2(\boldsymbol{\theta})\| &\le \gamma X_1(t;\beta) + \gamma^{-1}\sigma_t^2(\boldsymbol{\theta}) \\
\|\nabla\sigma_t(\boldsymbol{\theta})\| &\le \frac{1}{2}\cdot[\gamma^{1/2}\beta^{-1/2}X_2(t;\beta) + \gamma^{-1}\sigma_t(\boldsymbol{\theta})] \\
\|\nabla^2\sigma_t^2(\boldsymbol{\theta})\| &\le 2[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t;\beta) + X_1(t;\beta)] \\
\|\nabla^2\sigma_t(\boldsymbol{\theta})\| &\le \frac{\omega_0^{-1/2}}{4}\cdot\gamma\beta^{-1}X_2(t;\beta)^2 + \gamma^{-1/2}\beta^{-1/2}X_2(t;\beta) + \frac{1}{4}\gamma^{-2}\sigma_t(\boldsymbol{\theta}) \\
&\quad + \omega_0^{-1/2}\cdot[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t;\beta) + X_1(t;\beta)]
\end{aligned}$$

Proof of Lemma 8.

$$\sigma_t^2(\boldsymbol{\theta}) = \omega_0 + \beta\sigma_{t-1}^2(\boldsymbol{\theta}) + \gamma Y_{t-1}^2 = \frac{\omega_0}{1-\beta} + \gamma\sum_{i=1}^{\infty}\beta^{i-1}Y_{t-i}^2. \tag{73}$$

Therefore,

$$\begin{aligned}
\nabla\sigma_t^2(\boldsymbol{\theta}) &= \left[\omega_0/(1-\beta)^2 + \gamma\sum_{i=2}^{\infty}(i-1)\beta^{i-2}Y_{t-i}^2,\ \sum_{i=1}^{\infty}\beta^{i-1}Y_{t-i}^2, 0, 0\right] \\
&= \left[\omega_0/(1-\beta)^2 + \gamma X_1(t;\beta),\ \gamma^{-1}(\sigma_t^2 - \omega_0/(1-\beta)), 0, 0\right] \qquad (74) \\
\|\nabla\sigma_t^2(\boldsymbol{\theta})\| &\le \omega_0/(1-\beta)^2 + \gamma X_1(t;\beta) + \gamma^{-1}(\sigma_t^2 - \omega_0/(1-\beta)) \\
&= \gamma X_1(t;\beta) + \gamma^{-1}\sigma_t^2 + \frac{\omega_0}{1-\beta}\left(\frac{1}{1-\beta} - \frac{1}{\gamma}\right) \\
&\le \gamma X_1(t;\beta) + \gamma^{-1}\sigma_t^2,
\end{aligned}$$

since $\beta + \gamma \le 1$ implies $1/(1-\beta) - 1/\gamma \le 0$. Furthermore,

$$\|\nabla\sigma_t(\boldsymbol{\theta})\| = \frac{\|\nabla\sigma_t^2(\boldsymbol{\theta})\|}{2\sigma_t(\boldsymbol{\theta})} \le \frac{1}{2}\left[\frac{\gamma\sum_{i=2}^{\infty}(i-1)\beta^{i-2}Y_{t-i}^2}{\sqrt{\frac{\omega_0}{1-\beta} + \gamma\sum_{i=1}^{\infty}\beta^{i-1}Y_{t-i}^2}} + \frac{\gamma^{-1}\sigma_t^2}{\sigma_t}\right].$$

For all $j \geq 2$, we have

$$\frac{\gamma(j-1)\beta^{j-2}Y_{t-j}^2}{\sqrt{\frac{\omega_0}{1-\beta} + \gamma\sum_{i=1}^{\infty}\beta^{i-1}Y_{t-i}^2}} \leq \frac{\gamma(j-1)\beta^{j-2}Y_{t-j}^2}{\sqrt{\gamma\beta^{j-1}Y_{t-j}^2}} = (j-1)\gamma^{1/2}\beta^{(j-3)/2}|Y_{t-j}|$$

as all summands in the denominator are positive. This implies

$$\begin{aligned}\|\nabla\sigma_t(\boldsymbol{\theta})\| &\leq \frac{1}{2}\cdot\left[\gamma^{1/2}\beta^{-1/2}\sum_{i=2}^{\infty}(i-1)\beta^{(i-2)/2}|Y_{t-i}| + \gamma^{-1}\sigma_t(\boldsymbol{\theta})\right] \\ &= \frac{1}{2}\cdot[\gamma^{1/2}\beta^{-1/2}X_2(t;\beta) + \gamma^{-1}\sigma_t(\boldsymbol{\theta})].\end{aligned}$$

Using equation (74), we obtain

$$\nabla^2\sigma_t^2(\boldsymbol{\theta}) = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where

$$\begin{aligned}a_{11} &= \frac{2\omega_0}{(1-\beta)^3} + \gamma\sum_{i=3}^{\infty}(i-1)(i-2)\beta^{i-3}Y_{t-i}^2 \\ a_{12} = a_{21} &= \sum_{i=2}^{\infty}(i-1)\beta^{i-2}Y_{t-i}^2\end{aligned}$$

Thus, since the Frobenius norm is always less than the sum of the absolute values of the matrix entries,

$$\|\nabla^2\sigma_t^2(\boldsymbol{\theta})\| \leq 2[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t;\beta) + X_1(t;\beta)].$$

Using that $\nabla\sigma_t(\boldsymbol{\theta}) = \nabla\sigma_t^2(\boldsymbol{\theta})/(2\sigma_t(\boldsymbol{\theta}))$, we find that

$$\nabla^2\sigma_t(\boldsymbol{\theta}) = \frac{\nabla_t'^2(\boldsymbol{\theta})\nabla\sigma_t(\boldsymbol{\theta})}{-2\sigma_t^2(\boldsymbol{\theta})} + \frac{\nabla^2\sigma_t^2(\boldsymbol{\theta})}{2\sigma_t(\boldsymbol{\theta})} = \frac{\nabla'\sigma_t(\boldsymbol{\theta})\nabla\sigma_t(\boldsymbol{\theta})}{-\sigma_t(\boldsymbol{\theta})} + \frac{\nabla^2\sigma_t^2(\boldsymbol{\theta})}{2\sigma_t(\boldsymbol{\theta})},$$

therefore,

$$\|\nabla^2\sigma_t(\boldsymbol{\theta})\| \leq \frac{\|\nabla\sigma_t(\boldsymbol{\theta})\|^2}{\sigma_t(\boldsymbol{\theta})} + \frac{\|\nabla^2\sigma_t^2(\boldsymbol{\theta})\|}{2\sigma_t(\boldsymbol{\theta})}$$

Since $\sigma_t^2 \geq \omega_0 > 0$ and using our previous results, we obtain the claimed bound on $\|\nabla^2\sigma_t(\boldsymbol{\theta})\|$. ■

Lemma 9 *Under GARCH Assumption 2, it holds that*

$$\begin{aligned}
|v_t(\boldsymbol{\theta})| \leq & V(\mathcal{F}_{t-1}) = B_1 \cdot S_1(\mathcal{F}_{t-1}) \\
\|\nabla e_t(\boldsymbol{\theta})\| \leq & H_1(\mathcal{F}_{t-1}) = B_1 \cdot S_2(\mathcal{F}_{t-1}) + S_1(\mathcal{F}_{t-1}) \\
\|\nabla v_t(\boldsymbol{\theta})\| \leq & V_1(\mathcal{F}_{t-1}) = H_1(\mathcal{F}_{t-1}) + V(\mathcal{F}_{t-1}) \\
\|\nabla^2 e_t(\boldsymbol{\theta})\| \leq & H_2(\mathcal{F}_{t-1}) = B_1 \cdot S_3(\mathcal{F}_{t-1}) + 2S_2(\mathcal{F}_{t-1}) \\
\|\nabla^2 v_t(\boldsymbol{\theta})\| \leq & V_2(\mathcal{F}_{t-1}) = H_2(\mathcal{F}_{t-1}) + 2H_1(\mathcal{F}_{t-1}),
\end{aligned}$$

where

$$\begin{aligned}
S_1(\mathcal{F}_{t-1}) &= \sqrt{\omega_0 \delta_1^{-1} + \gamma \sum_{i=1}^{\infty} (1-\delta_1)^{i-1} Y_{t-i}^2} \\
S_2(\mathcal{F}_{t-1}) &= \frac{1}{2} \cdot [(1-\delta_1)^{1/2} \delta_1^{-1/2} X_2(t; 1-\delta_1) + \delta_1^{-1} S_1(\mathcal{F}_{t-1})] \\
S_3(\mathcal{F}_{t-1}) &= \frac{\omega_0^{-1/2}}{4} \cdot (\delta_1^{-1} - 1) X_2(t; 1-\delta_1)^2 + \delta_1^{-1} X_2(t; 1-\delta_1) + \frac{1}{4} \delta_1^{-2} S_1(\mathcal{F}_{t-1}) \\
&\quad + \omega_0^{-1/2} \cdot [\omega_0 \delta_1^{-3} + (1-\delta_1) X_3(t; 1-\delta_1) + X_1(t; 1-\delta_1)].
\end{aligned}$$

Proof of Lemma 9. As a function in β and γ, $\sigma_t(\boldsymbol{\theta})$ is increasing in both arguments, see equation (73), and, in fact, it does not depend on the parameters b and c. Therefore, $\sigma_t(\boldsymbol{\theta}) \leq S_1(\mathcal{F}_{t-1})$. The quantities $X_1(t,\beta)$, $X_2(t,\beta)$, $X_3(t,\beta)$ defined in the beginning of this section are all increasing in β, and thus, bounded by $X_1(t, 1-\delta_1)$, $X_2(t, 1-\delta_1)$, $X_3(t, 1-\delta_1)$, respectively. Recall that $|b| \leq B_1$ under GARCH Assumption 2.

The first inequality holds because $|v_t(\boldsymbol{\theta})| \leq |e_t(\boldsymbol{\theta})| = |b| \cdot \sigma_t(\boldsymbol{\theta})$. The remaining ones are implied by Lemma 8 and

$$\begin{aligned}
\|\nabla e_t(\boldsymbol{\theta})\| &= \|b \cdot \nabla \sigma_t(\boldsymbol{\theta}) + \sigma_t(\boldsymbol{\theta}) \cdot [0,0,1,0]\| \leq |b| \cdot \|\nabla \sigma_t(\boldsymbol{\theta})\| + \sigma_t(\boldsymbol{\theta}) \\
\|\nabla v_t(\boldsymbol{\theta})\| &= \|c \cdot \nabla e_t(\boldsymbol{\theta}) + e_t(\boldsymbol{\theta}) \cdot [0,0,0,1]\| \leq \|\nabla e_t(\boldsymbol{\theta})\| + |b| \cdot \sigma_t(\boldsymbol{\theta}) \\
\|\nabla^2 e_t(\boldsymbol{\theta})\| &= \|b \nabla^2 \sigma_t(\boldsymbol{\theta}) + [0,0,1,0]' \nabla \sigma_t(\boldsymbol{\theta}) + \nabla' \sigma_t(\boldsymbol{\theta}) [0,0,1,0]\| \\
&\leq |b| \cdot \|\nabla^2 \sigma_t(\boldsymbol{\theta})\| + 2\|\nabla \sigma_t(\boldsymbol{\theta})\| \\
\|\nabla^2 v_t(\boldsymbol{\theta})\| &= \|c \nabla^2 e_t(\boldsymbol{\theta}) + [0,0,0,1]' \nabla e_t(\boldsymbol{\theta}) + \nabla' e_t(\boldsymbol{\theta}) [0,0,0,1]\| \\
&\leq \|\nabla^2 e_t(\boldsymbol{\theta})\| + 2\|\nabla e_t(\boldsymbol{\theta})\|.
\end{aligned}$$

■

Lemma 10 *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of random variables and define $X = \sum_{i=1}^{\infty} a_i|X_i|$, where $a_i > 0$ for all $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty} a_i < \infty$. Let $p > 1$. If $\sup_{i\in\mathbb{N}} \mathbb{E}|X_i|^p \leq K < \infty$ for some constant K, then $\mathbb{E}|X|^p \leq (\sum_{i=1}^{\infty} a_i)^p K$.*

Proof of Lemma 10. By Jensen's inequality, $\mathbb{E}|Z|^p \geq |\mathbb{E}Z|^p$. We rewrite X as

$$X = \sum_{i=1}^{\infty} a_i|X_i| = \left(\sum_{i=1}^{\infty} a_i\right) \cdot \sum_{i=1}^{\infty} \left(\sum_{i=1}^{\infty} a_i\right)^{-1} a_i|X_i|.$$

Note that $\sum_{i=1}^{\infty}(\sum_{i=1}^{\infty} a_i)^{-1}a_i = 1$, namely $\{(\sum_{i=1}^{\infty} a_i)^{-1}a_i\}_{i=1}^{\infty}$ is a probability measure. Then, using Jensen's inequality,

$$X^p = \left(\sum_{j=1}^{\infty} a_j\right)^p \cdot \left(\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j\right)^{-1} a_i|X_i|\right)^p \leq \left(\sum_{j=1}^{\infty} a_j\right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j\right)^{-1} a_i|X_i|^p.$$

Thus,

$$\mathbb{E}\left[X^p\right] \leq \left(\sum_{j=1}^{\infty} a_j\right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j\right)^{-1} a_i\mathbb{E}|X_i|^p \leq \left(\sum_{j=1}^{\infty} a_j\right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j\right)^{-1} a_i K = \left(\sum_{j=1}^{\infty} a_j\right)^p K$$

■

Lemma 11 *Under GARCH Assumption 2 and for any $p > 1$, $p_1, \ldots, p_6 > 0$, the following statements hold for all t:*

(1) If $\mathbb{E}|Y_t|^p < \infty$, then the following quantities are all finite: $\mathbb{E}[V^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[V_1^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_1^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[V_2^{p/2}(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_2^{p/2}(\mathcal{F}_{t-1})]$.

(2) If $p = p_1 + p_2 + p_3 + 2p_4 + 2p_5 + p_6$ and $\mathbb{E}|Y_t|^p < \infty$, then

$$\mathbb{E}[V^{p_1}(\mathcal{F}_{t-1})V_1^{p_2}(\mathcal{F}_{t-1})H_1^{p_3}(\mathcal{F}_{t-1})V_2^{p_4}(\mathcal{F}_{t-1})H_2^{p_5}(\mathcal{F}_{t-1})|Y_t|^{p_6}] < \infty.$$

(3) If $\mathbb{E}|Y_t|^{4+\delta} < \infty$ for some $\delta > 0$, all the moment conditions in Assumption 2(D)' could be satisfied.

Proof of Lemma 11. Part (1) follows by combining Lemma 9 with Lemma 10, and part (2) is a consequence of part (1) and Hölder's inequality. ■

Lemma 11 implies that GARCH Assumption 3 implies Assumption 2(D) of the paper.

Appendix SA.2.6: Assumption 2(E)

$\mathbf{D}_0$ is the Hessian of the expected loss at $\boldsymbol{\theta}^0$, so it is positive semi-definite. Let $x = (x_1, \ldots, x_4)' \in \mathbb{R}^4$ such that $x'\mathbf{D}_0 x = 0$. This implies that $x'\nabla v_t(\boldsymbol{\theta}^0) = 0$, $x'\nabla e_t(\boldsymbol{\theta}^0) = 0$ almost surely. We have, $x'\nabla v_t(\boldsymbol{\theta}^0) = cx'\nabla e_t(\boldsymbol{\theta}^0) + x_4 e_t(\boldsymbol{\theta}^0)$. Therefore, $x_4 = 0$. Furthermore,

$$\begin{aligned} 2\sigma_t(\boldsymbol{\theta}^0)x'\nabla e_t(\boldsymbol{\theta}^0) &= 2\sigma_t(\boldsymbol{\theta}^0)bx'\nabla\sigma_t(\boldsymbol{\theta}^0) + 2x_3\sigma_t^2(\boldsymbol{\theta}^0) \\ &= bx'\nabla\sigma_t^2(\boldsymbol{\theta}^0) + 2x_3\sigma_t^2(\boldsymbol{\theta}^0) = 0, \text{ a.s.} \end{aligned} \tag{75}$$

The stationarity of $\{Y_t\}$ implies that $\sigma_t^2(\boldsymbol{\theta}^0)$ is stationary. Therefore it also holds that

$$bx'\nabla\sigma_{t-1}^2(\boldsymbol{\theta}^0) + 2x_3\sigma_{t-1}^2(\boldsymbol{\theta}^0) = 0, \text{ a.s.} \tag{76}$$

Computing (75) $-\beta\cdot$(76), we obtain that a.s.

$$0 = bx'[\sigma_{t-1}^2(\boldsymbol{\theta}^0), Y_{t-1}^2, 0, 0]' + 2x_3(\omega_0 + \gamma Y_{t-1}^2) = (bx_2 + 2\gamma x_3)Y_{t-1}^2 + (2\omega_0 x_3 + bx_1\sigma_{t-1}^2(\boldsymbol{\theta}^0)). \tag{77}$$

By the assumption that $Y_{t-1}|\sigma_{t-1}^2 \sim F_\eta(0, \sigma_{t-1}^2(\boldsymbol{\theta}^0))$ and that $\sigma_{t-1}^2(\boldsymbol{\theta}^0) = \omega_0 + \beta_0\sigma_{t-2}^2(\boldsymbol{\theta}^0) + \gamma_0 Y_{t-2}^2$, we can conclude from the above equation that $x_1 = x_2 = x_3 = 0$. Thus $\mathbf{D}_0$ is positive definite.

Appendix SA.2.7: Assumption 2(G)

We now verify this assumption for the GARCH(1,1) model. Set $a = bc$, so that $v_t = a\sigma_t$. Then for $T \geq 5$, a necessary condition for $Y_t = v_t(\boldsymbol{\theta})$, $t = 1, \ldots, T$ is given by the set of equations

$$Y_t^2 = a^2\beta^t\sigma_0^2 + a^2\beta^{t-1}(\omega_0 + \gamma Y_0^2) + a^2\sum_{k=1}^{t-1}\beta^{t-1-k}(\omega_0 + \gamma Y_k^2), \quad t = 1, \ldots, 4$$

or, equivalently,

$$Y_1^2 = a^2\beta\sigma_0^2 + a^2(\omega_0 + \gamma Y_0^2) \tag{78}$$

$$Y_2^2 = \beta Y_1^2 + a^2(\omega_0 + \gamma Y_1^2) \tag{79}$$

$$Y_3^2 = \beta Y_2^2 + a^2(\omega_0 + \gamma Y_2^2) \tag{80}$$

$$Y_4^2 = \beta Y_3^2 + a^2(\omega_0 + \gamma Y_3^2). \tag{81}$$

Solving equations (79)-(81) for β and equating the results, we obtain

$$\frac{a^2}{\omega_0} = \frac{Y_2^4 - Y_1^2Y_3^2}{Y_2^2 - Y_1^2} = \frac{Y_3^4 - Y_2^2Y_4^2}{Y_3^2 - Y_2^2}. \tag{82}$$

Therefore, a necessary condition such that $Y_t = v_t(\theta)$, $t = 1, \ldots, T$ for some parameter $\boldsymbol{\theta} \in \Theta$ is that $(Y_1, \ldots, Y_T)$ lies in the set $p^{-1}(0) = \{(Y_1, ..., Y_T) \in \mathbb{R}^T | p(Y_1, ..., Y_T) = 0\}$, where p is the polynomial function

$$p(Y_1, \ldots, Y_T) = (Y_2^4 - Y_1^2 Y_3^2)(Y_3^2 - Y_2^2) - (Y_3^4 - Y_2^2 Y_4^2)(Y_2^2 - Y_1^2).$$

The set $p^{-1}(0)$ has Hausdorff dimension less than T. Therefore, as the distribution of $(Y_1, \ldots, Y_T)$ is assumed to be absolutely continuous from GARCH Assumption 1, we obtain the claim with $K = 4$.

Appendix SA.2.8: Summary

We summarize the arguments showing that Assumption 1 and 2 of the paper are satisfied under GARCH Assumptions 1–3.

Assumption 1: Part (A) holds as it has been shown in Section SA.2.4.1 that the uniform law of large number holds under our GARCH Assumptions. Part (B)(i)-(ii) are satisfied under GARCH Assumptions 1-2. Part (B)(iii) is easy to check. Concerning Part (B)(iv), we have shown in Section SA.2.3 that the GARCH model is identifiable when ω is normalized.

Assumption 2: Part (A)(i) is easy to check, (ii) is satisfied by GARCH Assumption 1. Part (B)(i) is satisfied by GARCH Assumption 1, (ii) is clearly weaker than GARCH Assumption 3. Part (C)(i) follows easily from $\sigma_t(\boldsymbol{\theta})^2 \geq \omega_0 > 0$ and the bounds on the parameter $|b|$. Part (C)(ii) has been shown in Lemma 9. Part (D) is implied by Lemma 11. Part (E) is discussed in Section SA.2.6. Part (F) is satisfied under GARCH Assumptions 2–3 as discussed in Section SA.2.2, and Part (G) is satisfied by GARCH Assumption 1 as discussed in Section SA.2.7.

Appendix SA.3: Additional tables

Table S1: Finite-sample performance of (Q)MLE

	$T = 2500$			$T = 5000$		
	ω	β	γ	ω	β	γ
Panel A: N(0,1) innovations						
True	0.050	0.950	0.050	0.050	0.950	0.050
Median	0.053	0.897	0.050	0.051	0.899	0.050
Avg bias	0.011	(0.011)	0.000	0.005	(0.005)	0.000
St dev	0.056	0.064	0.013	0.023	0.029	0.009
Coverage	0.936	0.930	0.928	0.936	0.933	0.937
Panel B: Skew t (5,-0.5) innovations						
True	0.050	0.950	0.050	0.050	0.950	0.050
Median	0.052	0.895	0.049	0.052	0.897	0.050
Avg bias	0.017	(0.023)	0.005	0.006	(0.008)	0.002
St dev	0.077	0.095	0.028	0.026	0.037	0.017
Coverage	0.899	0.907	0.897	0.913	0.907	0.903

Notes: This table presents results from 1000 replications of the estimation of the parameters of a GARCH(1,1) model, using the Normal likelihood. In Panel A the innovations are standard Normal, and so estimation is then ML. In Panel B the innovations are standardized skew t, and so estimation is QML. Details are described in Section 4 of the main paper. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table S2: Simulation results for Normal innovations, estimation by CAViaR

	$T = 2500$			$T = 5000$		
	β	γ	a_α	β	γ	a_α
			$\alpha = 0.01$			
True	0.900	0.050	-2.326	0.900	0.050	-2.326
Median	0.901	0.048	-2.275	0.899	0.048	-2.347
Avg bias	-0.017	0.012	-0.120	-0.011	0.006	-0.095
St dev	0.079	0.066	0.957	0.051	0.034	0.718
Coverage	0.881	0.874	0.907	0.892	0.886	0.905
			$\alpha = 0.025$			
True	0.900	0.050	-1.960	0.900	0.050	-1.960
Median	0.898	0.047	-1.953	0.896	0.047	-2.009
Avg bias	-0.018	0.005	-0.136	-0.012	0.002	-0.110
St dev	0.068	0.038	0.728	0.052	0.023	0.566
Coverage	0.906	0.879	0.934	0.913	0.892	0.918
			$\alpha = 0.05$			
True	0.900	0.050	-1.645	0.900	0.050	-1.645
Median	0.901	0.047	-1.639	0.899	0.049	-1.667
Avg bias	-0.014	0.005	-0.085	-0.009	0.002	-0.070
St dev	0.068	0.037	0.597	0.045	0.023	0.436
Coverage	0.909	0.884	0.930	0.918	0.900	0.935
			$\alpha = 0.10$			
True	0.900	0.050	-1.282	0.900	0.050	-1.282
Median	0.898	0.047	-1.291	0.898	0.048	-1.289
Avg bias	-0.016	0.006	-0.076	-0.010	0.003	-0.055
St dev	0.069	0.041	0.482	0.047	0.025	0.364
Coverage	0.916	0.883	0.933	0.921	0.896	0.937
			$\alpha = 0.20$			
True	0.900	0.050	-0.842	0.900	0.050	-0.842
Median	0.898	0.048	-0.848	0.899	0.048	-0.840
Avg bias	-0.023	0.022	-0.058	-0.016	0.007	-0.049
St dev	0.091	0.107	0.391	0.063	0.044	0.304
Coverage	0.914	0.876	0.931	0.929	0.901	0.940

Notes: This table presents results from 1000 replications of the estimation of VaR from a GARCH(1,1) DGP with standard Normal innovations. Details are described in Section 4 of the main paper. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table S3: Simulation results for skew t innovations, estimation by CAViaR

	$T = 2500$			$T = 5000$		
	β	γ	a_α	β	γ	a_α
			$\alpha = 0.01$			
True	0.900	0.050	-3.290	0.900	0.050	-3.290
Median	0.898	0.045	-3.272	0.899	0.045	-3.306
Avg bias	-0.041	0.022	-0.355	-0.027	0.008	-0.306
St dev	0.142	0.097	1.928	0.103	0.044	1.546
Coverage	0.771	0.805	0.827	0.785	0.808	0.823
			$\alpha = 0.025$			
True	0.900	0.050	-2.408	0.900	0.050	-2.408
Median	0.899	0.047	-2.371	0.898	0.049	-2.414
Avg bias	-0.026	0.012	-0.190	-0.016	0.004	-0.144
St dev	0.103	0.067	1.135	0.070	0.033	0.862
Coverage	0.832	0.841	0.877	0.830	0.862	0.859
			$\alpha = 0.05$			
True	0.900	0.050	-1.800	0.900	0.050	-1.800
Median	0.899	0.047	-1.780	0.899	0.049	-1.792
Avg bias	-0.023	0.008	-0.146	-0.013	0.004	-0.087
St dev	0.092	0.060	0.782	0.057	0.028	0.563
Coverage	0.863	0.861	0.892	0.883	0.871	0.890
			$\alpha = 0.10$			
True	0.900	0.050	-1.223	0.900	0.050	-1.223
Median	0.900	0.049	-1.205	0.900	0.049	-1.217
Avg bias	-0.019	0.008	-0.074	-0.010	0.004	-0.043
St dev	0.080	0.050	0.495	0.050	0.027	0.356
Coverage	0.895	0.892	0.919	0.892	0.905	0.910
			$\alpha = 0.20$			
True	0.900	0.050	-0.652	0.900	0.050	-0.652
Median	0.903	0.051	-0.619	0.902	0.051	-0.636
Avg bias	-0.027	0.026	-0.035	-0.016	0.009	-0.028
St dev	0.122	0.109	0.353	0.084	0.042	0.271
Coverage	0.867	0.887	0.897	0.890	0.889	0.916

Notes: This table presents results from 1000 replications of the estimation of VaR from a GARCH(1,1) DGP with skew t innovations. Details are described in Section 4 of the main paper. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table S4: Diebold-Mariano t-statistics on average out-of-sample loss differences for the DJIA, NIKKEI and FTSE100 (alpha=0.05)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
					Panel A: DJIA					
RW125		-2.200	-3.536	2.324	2.860	2.935	3.006	3.821	3.244	3.494
RW250	2.200		-3.349	2.983	3.411	3.502	3.989	4.522	3.926	3.957
RW500	3.536	3.349		3.979	4.336	4.417	4.805	5.321	4.829	4.860
G-N	-2.324	-2.983	-3.979		3.573	2.787	0.791	1.419	1.472	1.670
G-Skt	-2.860	-3.411	-4.336	-3.573		1.385	-0.034	0.625	0.195	0.302
G-EDF	-2.935	-3.502	-4.417	-2.787	-1.385		-0.266	0.432	-0.119	-0.031
FZ-2F	-3.006	-3.989	-4.805	-0.791	0.034	0.266		1.085	0.192	0.247
FZ-1F	-3.821	-4.522	-5.321	-1.419	-0.625	-0.432	-1.085		-0.796	-0.613
G-FZ	-3.244	-3.926	-4.829	-1.472	-0.195	0.119	-0.192	0.796		0.126
Hybrid	-3.494	-3.957	-4.86	-1.670	-0.302	0.031	-0.247	0.613	-0.126	
					Panel B: NIKKEI					
RW125		-0.225	-1.047	3.703	3.687	3.719	3.733	3.219	3.692	3.868
RW250	0.225		-1.162	4.048	4.058	4.098	3.897	3.582	4.076	4.249
RW500	1.047	1.162		3.733	3.748	3.785	3.768	3.387	3.773	3.847
G-N	-3.703	-4.048	-3.733		1.165	2.110	-1.841	-1.261	1.861	0.457
G-Skt	-3.687	-4.058	-3.748	-1.165		1.797	-1.888	-1.378	1.468	0.295
G-EDF	-3.719	-4.098	-3.785	-2.110	-1.797		-1.984	-1.522	-0.797	0.100
FZ-2F	-3.733	-3.897	-3.768	1.841	1.888	1.984		1.209	1.958	2.489
FZ-1F	-3.219	-3.582	-3.387	1.261	1.378	1.522	-1.209		1.487	2.624
G-FZ	-3.692	-4.076	-3.773	-1.861	-1.468	0.797	-1.958	-1.487		0.134
Hybrid	-3.868	-4.249	-3.847	-0.457	-0.295	-0.100	-2.489	-2.624	-0.134	

Table continued on next page.

Table S4 (cont'd): Diebold-Mariano t-statistics on average out-of-sample loss differences for the DJIA, NIKKEI and FTSE100 (alpha=0.05)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
					Panel C: FTSE					
RW125		-2.329	-3.439	3.275	3.485	3.450	2.732	3.279	3.300	3.141
RW250	2.329		-2.751	4.146	4.337	4.305	3.663	4.264	4.160	4.025
RW500	3.439	2.751		4.682	4.845	4.817	4.232	4.848	4.696	4.661
G-N	-3.275	-4.146	-4.682		4.327	4.446	-0.210	-0.070	0.581	1.048
G-Skt	-3.485	-4.337	-4.845	-4.327		-3.853	-0.746	-0.877	-4.066	0.428
G-EDF	-3.450	-4.305	-4.817	-4.446	3.853		-0.648	-0.731	-3.949	0.545
FZ-2F	-2.732	-3.663	-4.232	0.210	0.746	0.648		0.213	0.249	1.401
FZ-1F	-3.279	-4.264	-4.848	0.070	0.877	0.731	-0.213		0.128	1.321
G-FZ	-3.300	-4.160	-4.696	-0.581	4.066	3.949	-0.249	-0.128		1.006
Hybrid	-3.141	-4.025	-4.661	-1.048	-0.428	-0.545	-1.401	-1.321	-1.006	

Notes: This table presents t-statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2 of the main paper.

Table S5: Out-of-sample average losses and goodness-of-fit tests (alpha=0.025)

	Average loss				*GoF p-values: VaR*				*GoF p-values: ES*			
	S&P	**DJIA**	**NIK**	**FTSE**	**S&P**	**DJIA**	**NIK**	**FTSE**	**S&P**	**DJIA**	**NIK**	**FTSE**
RW-125	1.119	1.088	1.525	1.166	0.036	0.004	0.001	0.000	0.017	0.006	0.001	0.001
RW-250	1.164	1.117	1.525	1.209	0.009	0.009	0.006	0.000	0.037	*0.056*	0.015	0.006
RW-500	1.245	1.187	1.561	1.294	0.003	0.001	0.011	0.000	0.032	0.025	0.014	0.000
GCH-N	1.089	1.021	1.341	1.052	0.000	0.001	**0.177**	0.000	0.000	0.000	*0.053*	0.000
GCH-Skt	1.044	0.978	**1.328**	**1.026**	0.008	0.009	**0.796**	0.001	0.011	0.006	**0.725**	0.001
GCH-EDF	*1.028*	*0.969*	1.329	*1.042*	**0.188**	0.031	**0.796**	0.000	**0.258**	0.017	**0.593**	0.000
FZ-2F	1.039	0.998	1.421	1.242	0.000	0.002	**0.341**	0.000	0.001	0.001	**0.158**	0.000
FZ-1F	1.030	0.985	1.390	1.056	*0.057*	0.007	**0.773**	0.000	**0.130**	*0.058*	**0.415**	0.000
GCH-FZ	**1.020**	**0.951**	**1.328**	1.055	**0.125**	**0.364**	**0.688**	0.000	*0.222*	*0.403*	**0.521**	0.000
Hybrid	1.053	1.030	1.345	1.079	0.001	**0.114**	**0.558**	0.000	0.002	*0.075*	**0.464**	0.000

Notes: The left panel of this table presents the average losses, using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The lowest average loss in each column is highlighted in bold, the second-lowest is highlighted in italics. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2 of the main paper. The middle and right panels of this table present p-values from goodness-of-fit tests of the VaR and ES forecasts respectively. Values that are greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and values between 0.05 and 0.10 are in italics.

Table S6: Diebold-Mariano t-statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 (alpha=0.025)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
Panel A: S&P 500										
RW125		-1.836	-2.988	1.025	2.479	2.788	2.146	3.371	2.891	2.419
RW250	1.836		-2.815	1.725	2.747	3.004	2.602	3.712	3.135	2.992
RW500	2.988	2.815		2.823	3.673	3.893	3.630	4.624	4.023	4.045
G-N	-1.025	-1.725	-2.823		4.019	3.368	2.083	2.429	3.698	1.928
G-Skt	-2.479	-2.747	-3.673	-4.019		2.275	0.270	0.815	2.742	-0.594
G-EDF	-2.788	-3.004	-3.893	-3.368	-2.275		-0.592	-0.074	1.393	-1.483
FZ-2F	-2.146	-2.602	-3.630	-2.083	-0.270	0.592		0.487	1.227	-0.729
FZ-1F	-3.371	-3.712	-4.624	-2.429	-0.815	0.074	-0.487		0.579	-1.605
G-FZ	-2.891	-3.135	-4.023	-3.698	-2.742	-1.393	-1.227	-0.579		-2.172
Hybrid	-2.419	-2.992	-4.045	-1.928	0.594	1.483	0.729	1.605	2.172	
Panel B: DJIA										
RW125		-0.971	-2.294	1.892	2.981	3.051	3.132	3.590	3.332	1.840
RW250	0.971		-2.527	1.954	2.844	2.968	3.640	3.732	3.311	2.043
RW500	2.294	2.527		2.891	3.717	3.852	4.680	4.679	4.195	3.093
G-N	-1.892	-1.954	-2.891		3.705	2.900	0.765	1.305	3.236	-0.459
G-Skt	-2.981	-2.844	-3.717	-3.705		1.421	-0.706	-0.291	2.335	-2.666
G-EDF	-3.051	-2.968	-3.852	-2.900	-1.421		-1.022	-0.705	2.213	-2.693
FZ-2F	-3.132	-3.640	-4.680	-0.765	0.706	1.022		1.344	1.740	-1.229
FZ-1F	-3.590	-3.732	-4.679	-1.305	0.291	0.705	-1.344		1.539	-1.943
G-FZ	-3.332	-3.311	-4.195	-3.236	-2.335	-2.213	-1.740	-1.539		-3.127
Hybrid	-1.840	-2.043	-3.093	0.459	2.666	2.693	1.229	1.943	3.127	

Table continued on next page.

Table S6 (cont'd): Diebold-Mariano t-statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 (alpha=0.025)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
					Panel C: NIKKEI					
RW125		0.010	-0.901	3.956	3.896	3.944	3.703	3.093	3.895	3.829
RW250	-0.010		-1.486	4.105	4.149	4.177	3.544	3.340	4.136	4.102
RW500	0.901	1.486		3.935	3.999	4.012	3.886	3.441	3.980	3.996
G-N	-3.956	-4.105	-3.935		1.799	2.032	-2.541	-2.010	2.052	-0.226
G-Skt	-3.896	-4.149	-3.999	-1.799		-0.785	-2.726	-2.532	-0.310	-0.977
G-EDF	-3.944	-4.177	-4.012	-2.032	0.785		-2.741	-2.499	0.459	-0.903
FZ-2F	-3.703	-3.544	-3.886	2.541	2.726	2.741		1.481	2.687	2.739
FZ-1F	-3.093	-3.34	-3.441	2.010	2.532	2.499	-1.481		2.454	2.971
G-FZ	-3.895	-4.136	-3.98	-2.052	0.310	-0.459	-2.687	-2.454		-0.919
Hybrid	-3.829	-4.102	-3.996	0.226	0.977	0.903	-2.739	-2.971	0.919	
					Panel D: FTSE					
RW125		-1.557	-3.197	2.938	3.467	3.157	-1.683	2.978	2.570	2.173
RW250	1.557		-2.864	3.646	4.172	3.863	-0.758	3.788	3.355	2.985
RW500	3.197	2.864		4.350	4.789	4.532	1.179	4.688	4.173	3.972
G-N	-2.938	-3.646	-4.350		4.520	3.634	-3.549	-0.239	-0.340	-2.352
G-Skt	-3.467	-4.172	-4.789	-4.520		-4.471	-3.863	-1.996	-3.05	-3.991
G-EDF	-3.157	-3.863	-4.532	-3.634	4.471		-3.686	-0.949	-1.612	-3.218
FZ-2F	1.683	0.758	-1.179	3.549	3.863	3.686		3.924	3.468	3.271
FZ-1F	-2.978	-3.788	-4.688	0.239	1.996	0.949	-3.924		0.046	-1.602
G-FZ	-2.570	-3.355	-4.173	0.340	3.050	1.612	-3.468	-0.046		-2.354
Hybrid	-2.173	-2.985	-3.972	2.352	3.991	3.218	-3.271	1.602	2.354	

Notes: This table presents t-statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 2 of the main paper.

References

[1] Carrasco, M. and X. Chen, 2002, Mixing and moment properties of various GARCH and stochastic volatility models, *Econometric Theory*, 18(1), 17-39.

[2] Davidson, J. 1994, *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

[3] Lumsdaine, R., 1996, Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models, *Econometrica*, 64, 575–596.